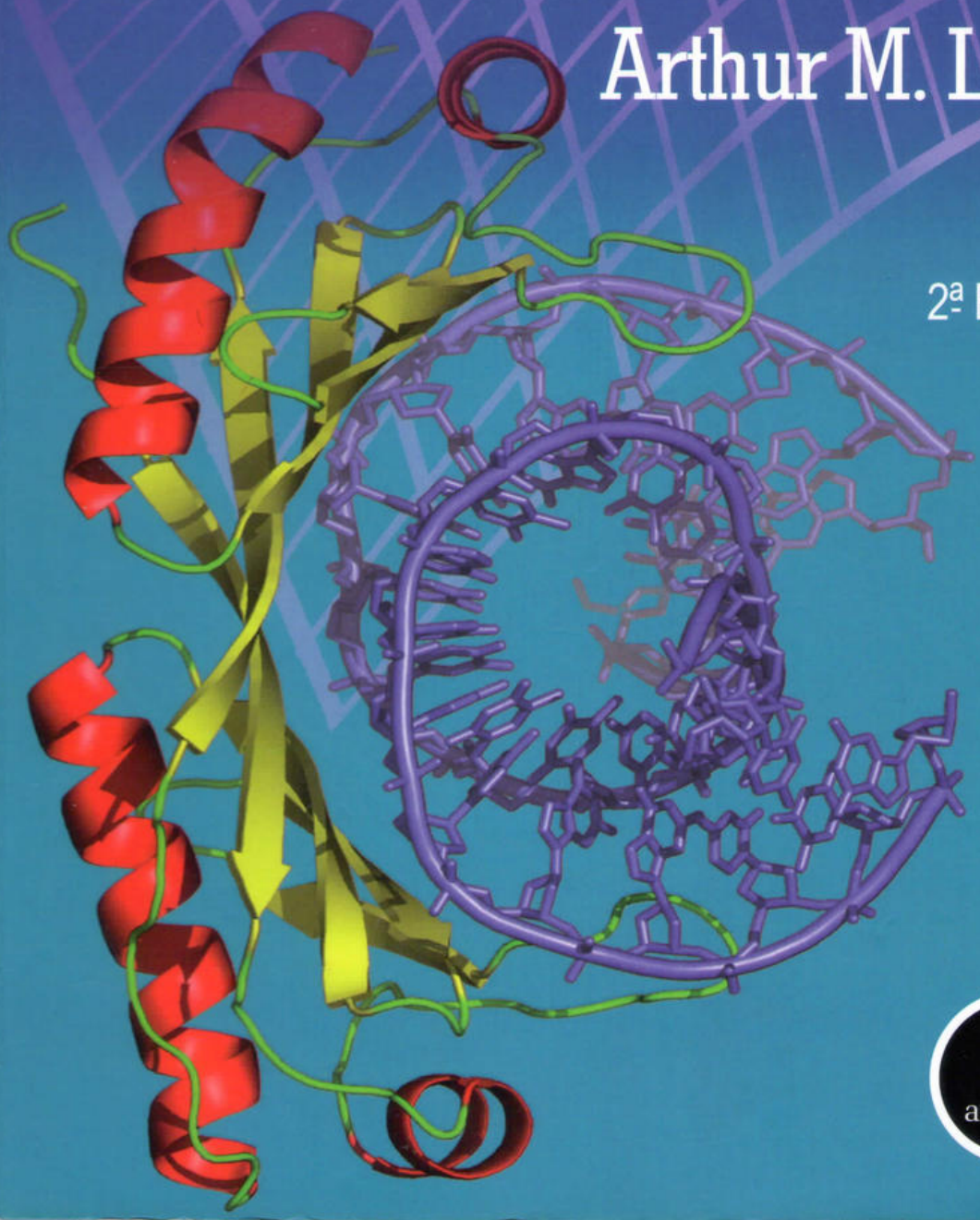
Introdução à Bioinformática

Arthur M. Lesk



2ª Edição

artmed®

Equipe de tradução

Ardala Elisa Breda Andrade
Mestre em Biologia Celular e Molecular, ênfase em Biofísica Molecular Computacional, pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS. Membro do LABIO/FACIN da PUCRS.

Evelyn Koeche Schroeder
Doutora em Biologia Celular e Molecular, ênfase em Biofísica Molecular Computacional, pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS. Membro do LABIO/FACIN da PUCRS.

Karina dos Santos Machado
Mestre em Ciência da Computação, ênfase em Bioinformática, pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS. Membro do LABIO/FACIN da PUCRS.

Osmar Norberto de Souza
Ph.D. em Biofísica Molecular Computacional pela Universidade de Londres, Londres, Inglaterra. Professor e Pesquisador da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) nas áreas de Bioinformática e Biofísica Molecular Computacional. Coordenador do LABIO/FACIN da PUCRS. Pesquisador do CNPq.

CÓPIA NÃO AUTORIZADA É CRIME
ABDR
ASSOCIAÇÃO BRASILEIRA DE DIREITOS REPROGRÁFICOS
RESPEITE O DIREITO AUTORAL

L629i Lesk, Arthur M.
Introdução à Bioinformática / Arthur M. Lesk ; tradução Ardala Elisa Breda Andrade, ... [et al.]. – 2. ed. – Porto Alegre : Artmed, 2008.
384 p. ; 25 cm.

ISBN 978-85-363-1104-3

1. Biologia. 2. Informática. I. Título.

573:004

Catalogação na publicação: Juliana Lagôas Coelho – CRB 10/1798

Arthur M. Lesk
Professor of Biochemistry and
Molecular Biology at The Pennsylvania State University, USA.

Introdução à Bioinformática

2ª Edição

Consultoria, supervisão e revisão técnica desta edição

Osmar Norberto de Souza
Ph.D. em Biofísica Molecular Computacional pela Universidade de Londres, Londres, Inglaterra.
Professor e Pesquisador da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) nas áreas de Bioinformática e Biofísica Molecular Computacional.
Coordenador do Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas (LABIO) da Faculdade de Informática (FACIN) da PUCRS.
Pesquisador do CNPq.

artmed®
2008

Obra originalmente publicada em língua inglesa sob o título
Introduction to Bioinformatics, 2nd Edition, ISBN 0-19-927787-7, © 2005. Editora original Oxford University Press. Esta edição não está a venda em Portugal.
Introduction to Bioinformatics, 2nd Edition, was originally published in English in 2005. This translation is published by arrangement with Oxford University Press. This edition is not for sale in Portugal.

© 2005, Arthur M. Lesk

Capa: *Mário Röhnelt*
Imagem: A ilustração representa a interação entre o domínio de ligação à caixa TATA (*TATA-box*) da proteína TBP (*TATA-box binding protein*) humana e um fragmento de DNA de 16 pares de base contendo a caixa TATA. A formação deste complexo é essencial para a transcrição de genes estruturais em eucariotos. TBP dobra a caixa TATA a uma curvatura de aproximadamente 83°. Apenas a cadeia principal, no formato *ribbons*, da proteína TBP está representada. Nela, as hélices α e as folhas β estão coloridas em vermelho e amarelo, respectivamente. As voltas e alças estão coloridas em verde. O DNA dupla fita está representado na forma de palitos, com o esqueleto fosfático destacado como uma tira cilíndrica, e com a fita 5'-3' colorida violeta escuro e a fita 3'-5' violeta claro. A estrutura tridimensional deste complexo [código PDB: 1ª, w] foi determinada por cristalografia por difração de raios X. [Fonte: Nikolov, D. B., Chen, H., Halay, E. D., Roeder, R. G., Burley, S. K. 1996. Crystal structure of a human TATA box-binding protein/TATA element complex, *Proc. Natl. Acad. Sci. U. S. A.* **93**:4862-4867.] Imagem gerada com o programa PyMol 0.95 [DeLano. W.L.: The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA (2002).]

Preparação do original: *Heloísa Stefan*

Leitura final: *Carla Bigliardi*

Supervisão editorial: *Letícia Bispo de Lima*

Editoração eletrônica: *Techbooks*

Reservados todos os direitos de publicação, em língua portuguesa, à
ARTMED® EDITORA S.A.
Av. Jerônimo de Ornelas, 670 – Santana
90040-340 – Porto Alegre – RS
Fone: (51) 3027-7000 Fax: (51) 3027-7070

É proibida a duplicação ou reprodução deste volume, no todo ou em parte, sob quaisquer formas ou por quaisquer meios (eletrônico, mecânico, gravação, fotocópia, distribuição na *Web* e outros), sem permissão expressa da Editora.

SÃO PAULO
Av. Angélica, 1.091 – Higienópolis
01227-100 – São Paulo – SP
Fone: (11) 3665-1100 Fax: (11) 3667-1333

SAC 0800 703-3444

IMPRESSO NO BRASIL
PRINTED IN BRAZIL

Daniella Costa

Este livro é dedicado à Eda, com quem eu combinei meus genes.

Do livro dos segredos infinitos da natureza
um pouco eu posso compreender.
– William Shakespeare em *Antony e Cleópatra*

Agradecimentos

Sou grato a muitos colegas por discussões e sugestões durante a preparação deste livro, às universidades de Uppsala, Umeå, Roma "Tor Vergata" e Cambrigde pela oportunidade de experimentar este material, e ao Institut des Haute Études Scientifiques, Bures-sur-Yvette, pela hospitalidade durante a escrita de parte deste livro.

Agradeço a D. J. Abraham, S. Aparicio, M. M. Babu, T. Baglin, D. Baker, A. Bench, G. Blatch, M. Brand, A. Brazma, A. Buckle, C. Cantor, R. W. Carrell, C. Chothia, A. Coulson, D. Crowther, T. Dafforn, R. Foley, A. Friday, M. B. Gerstein, T. Gibson, J. Irving, B. Jorden, J. Karn, K. Karplus, P. Klappa, E. V. Koonin, M. Krichevsky, P. Lawrence, A. Lister, E. L. Lesk, M. E. Lesk, V. E. Lesk, V. I. Lesk, M. Levitt, D. Liberles, O. Lichtarge, L. Lo Conte, D. A. Lomas, K. Lutfy, G. Magklaras, J. Magré, A. D. MacKerell, Jr., M. McFall-Ngai, J. McInerney, P. Miller, C. Mitchell, J. Moult, E. Nacheva, C. Notredame, H. Parfrey, D. Parkinson, A. Pastore, D. Penny, J. Pettitt, C. A. Praul, F. W. Roberts, G. D. Rose, P. B. Rosenthal, B. Rost, E. J. Simon, J. Sulston, M. Segal, O. Skovegaard, E. L. Sonnhammer, R. Srinivasan, R. Staden, I. Tickle, A. Tramontano, A. A. Travers, A. R. Venkitaraman, P. L. Welcsh, G. Vriend, J. C. Whisstock, M. Widersten, A. S. Wilkins, S. H. White e E. B. Ziff por sugestões e leitura crítica do manuscrito.

Agradeço à equipe que trabalhou na produção editorial desta obra por sua competência e paciência.

A. M. L.

Prefácio à 2ª Edição

A bioinformática cresceu desde que a 1ª edição deste livro surgiu.

A mudança mais surpreendente foi a refocalização na integração, isto é, tentar compreender os processos da vida como um sistema unificado. Eu escrevi, no final de *Introduction to Protein Science: Architecture, Function and Genomics*, "Durante o último século, biólogos moleculares investigaram os seres vivos em partes separadas. Nossa tarefa agora é colocá-las juntas novamente". Nós temos grandes quantidades de dados. Agora estamos tentando ver como eles se interrelacionam. No cerne dos processos da vida, encontram-se padrões complicados de interação entre os seus componentes, no tempo e no espaço. Para entender esses padrões, a área mudou no sentido de combinar informações em redes e tentar entender suas estruturas e dinâmica.

Dando suporte a este empreendimento está o crescimento da abundância de dados. O genoma humano, disponível na forma de um esboço quando a primeira versão apareceu, está completo agora. Juntam-se a ele os genomas completos e publicados de 608 organismos, mais os seqüenciamentos em andamento de genomas de 61 arqueobactérias, 1.297 bactérias, 773 eucariotos, 99 metagenomas e muitas outras seqüências de organelas e vírus. Esses genomas ajudam a compreender uns aos outros. Uma história que eles contam é sobre unidades fundamentais desconhecidas em todas as formas vivas, apesar das diferenças óbvias e profundas em morfologia e estilo de vida.

Seqüências genômicas são complementadas por outros tipos de dados, especialmente o proteoma. Padrões de expressão de genes e redes de interações reguladoras mostram como células e organismos implementam a informação no DNA. A potencialidade para a vida de um organismo está contida em seu genoma, mas seria impossível deduzir uma biografia a partir dele. Genomas não são fórmulas ou *scripts*. É nas proteínas, e em suas próprias interações e com o DNA, que devemos procurar pelo conjunto de atividades, dependendo de como respondem ao ambiente. A proteômica está nos fornecendo as informações de que precisamos para ver como os sistemas funcionam.

Pesquisas e aplicações necessitam que os dados estejam disponíveis em uma forma utilizável. Tornar os dados públicos não é suficiente. A informação deve ser submetida a um controle de qualidade, anotação, e uma estrutura lógica deve ser imposta para tornar essa informação recuperável. Por essas razões, estamos em débito com as instituições que armazenam, administram, organizam e distribuem os dados. Uma tendência recente parece ser a união desses grupos em projetos colaborativos que atravessam os continentes. De acordo com a necessidade de integrar o estudo de diferentes tipos de dados, estamos nos movendo em direção a um único repositório de dados biológicos. Cientistas individuais estarão aptos a definir "bancos de dados virtuais", adaptando o acesso às informações para suprir necessidades e interesses particulares.

Uma conseqüência gratificante para acadêmicos em bioinformática são as suas contribuições para aplicações em medicina, agricultura e tecnologia. Uma melhor compreensão dos processos da vida nos possibilitará lidar com eles quando falharem.

A.M.L.

Prefácio da 1ª Edição

Em 26 de junho de 2000, as ciências biológicas e médicas mudaram para sempre. O Primeiro-Ministro do Reino Unido, Tony Blair, e o Presidente dos Estados Unidos, Bill Clinton, realizaram uma conferência de imprensa comum, via satélite, para anunciar o término de um esboço da seqüência do Genoma Humano. O jornal *The New York Times* publicou como manchete: "Código Genético da Vida Humana é Decifrado pelos Cientistas". A seqüência de três bilhões de bases foi o resultado de mais de uma década de trabalho, durante a qual o objetivo sempre esteve visível e as únicas questões eram o quão rápido a tecnologia poderia progredir e quão generosos seriam os investimentos. O Quadro a seguir mostra alguns dos marcos ao longo deste caminho.

Os cientistas permaneceram lado a lado com os políticos. John Sulston, Diretor do The Sanger Centre no Reino Unido, foi uma peça-chave desde o início dos métodos de seqüenciamento rápido e em larga escala. Ele cresceu com o projeto desde o "primeiro participante e o seu treinamento" até a formação do consórcio internacional atual. Nos Estados Unidos, aparecendo juntamente com o Presidente Clinton, estavam Francis Collins, Diretor do Instituto de Pesquisa US National Human Genome, representando os esforços do financiamento público americano, e J. Graig Venter, Presidente e Diretor Científico da Celera Genomics Corporation, representando o setor comercial. É difícil introduzir estes dois pesquisadores sem pensar "neste lado... e neste lado..." Embora nunca efetivamente tenha havido uma luta, certamente existiu intensa competição e, nos últimos estágios, uma corrida.

A corrida foi mais do que um esforço para terminar primeiro e receber os créditos científicos por primazia. Na verdade, era uma corrida após a qual os concorrentes seriam testados não para saber se eles tinham tomado medicamentos, mas sim para saber se eles e outros poderiam descobri-los. Aplicações clínicas foram o principal motivo para o apoio dado ao projeto genoma humano. Uma vez decidido pelas cortes de justiça que seqüências de genes eram patenteáveis – com enorme potencial de ganho para fármacos baseados nessas seqüências – o setor comercial se apressou para submeter solicitações de patentes de conjunto de seqüências que eles determinaram, e os grupos acadêmicos correram para colocar cada pedaço de seqüência que *eles* determinaram no domínio público para *evitar* que Celera – ou qualquer outra companhia – solicitasse patentes.

Os grupos acadêmicos que se juntaram contra Celera eram de laboratórios colaboradores, sobretudo, mas não exclusivamente, no Reino Unido e nos Estados Unidos. Eles incluíam o Sanger Centre na Inglaterra, a Washington University em St. Louis, Missouri, o Whitehead Institute no Massachusetts Institute of Technology em Cambridge, Massachusetts, o Baylor College of Medicine em Houston, Texas, o Joint Genome Institute no Lawrence Livermore National Laboratory em Livermore, Califórnia, e o RIKEN Genomic Sciences Center, agora em Yokahama, Japão.

Ambos os lados podiam pagar suas próprias contas. Celera teve sua origem no capital de risco de sua matriz, a PE Corporation, e, após se tornar pública, de qualquer pessoa que se interessasse por uma agitação. O Sanger Centre era mantido pelo Medical Research Council do Reino Unido e pelo The Wellcome

Trust. Os laboratórios acadêmicos dos Estados Unidos eram mantidos pelo National Institutes of Health e pelo Department of Energy.

Em 26 de junho de 2000, os concorrentes concordaram em declarar a corrida empatada, ou pelo menos como uma foto de final de corrida desfocada, cuidadosamente manipulada.

Marcos no Projeto Genoma Humano

1953	A estrutura do DNA de Watson-Crick é publicada.
1975	F. Sanger e, independentemente, A. Maxam e W. Gilbert desenvolvem métodos para seqüenciamento de DNA.
1977	O bacteriófago ΦX-174 é seqüenciado: primeiro "genoma completo".
1980	A Suprema Corte dos Estados Unidos mantém que bactérias geneticamente modificadas são patenteáveis. Esta decisão foi a base original para o patenteamento de genes.
1981	O DNA mitocondrial humano é seqüenciado: 16.569 pares de bases.
1984	O genoma do vírus Epstein-Barr é seqüenciado: 172.281 pares de bases.
1990	É lançado o Projeto Internacional Genoma Humano – previsão de 15 anos para completá-lo.
1991	J. Craig Venter e colegas identificam genes ativos por meio de etiquetas de seqüência expressa (do inglês, EST, Expressed Sequence Tags) – seqüências de porções iniciais de DNA complementar ao RNA mensageiro.
1992	É completado o mapa de ligação de baixa resolução do genoma humano.
1992	Início do projeto de seqüenciamento do genoma de *Caenorhabditis elegans*.
1992	O Wellcome Trust e o Medical Research Council do Reino Unido estabelecem o Sanger Centre para seqüenciamento de genomas em larga escala, dirigido por J. Sulston.
1992	J. Craig Venter forma o The Institute for Genome Research (TIGR), associado com planos de explorar comercialmente o seqüenciamento por meio da identificação de genes e do descobrimento de novos fármacos.
1995	Término da primeira seqüência completa do genoma de uma bactéria, *Haemophilus influenzae*, pelo TIGR.
1996	Mapa de alta resolução do genoma humano – marcadores espaçados por ~600.000 pares de bases.
1996	Término do genoma da levedura, primeira seqüência completa de um genoma eucariótico.
Maio de 1998	Celera anuncia que pode terminar o genoma humano em 2001. O Wellcome Trust responde aumentando os investimentos no Sanger Centre.
1998	A seqüência do genoma de *Caenorhabditis elegans* é publicada.
1º de setembro de 1999	A seqüência do genoma de *Drosophila melanogaster* é anunciada por Celera Genomics e publicada na primavera de 2000.
1999	O Projeto Genoma Humano declara seu objetivo: trabalhar no esboço do genoma humano para 2001 (90% dos genes seqüenciados com mais que 95% de acurácia).
1º de dezembro de 1999	A seqüência do primeiro cromossomo humano completo é publicada.
26 de junho de 2000	Anúncio em conjunto do término do esboço da seqüência do genoma humano.
2003	50º aniversário da descoberta da estrutura do DNA. Término da seqüência do genoma humano de alta resolução feita por um consórcio público.

O genoma humano é somente uma das muitas seqüências de genomas completas conhecidas. Vistas como um todo, as seqüências de genomas de organismos distribuídos amplamente ao longo dos ramos da árvore da vida nos permitem uma percepção, apenas sugerida anteriormente, da enorme unidade de *detalhes* de toda a vida na Terra. Elas mudaram nossas percepções, da mesma forma que as primeiras imagens da Terra tiradas do espaço produziram uma visão unificada do nosso planeta.

O seqüenciamento do genoma humano é comparado com o Projeto Manhattan, que produziu as armas nucleares durante a Segunda Guerra Mundial, e com o programa espacial que enviou pessoas à Lua, como um dos maiores avanços em realizações tecnológicas do último século. Esses projetos compartilham fundamentos científicos básicos e o desenvolvimento e o apoio em larga escala e de alto custo da engenharia. Para a biologia, nem as atitudes nem os orçamentos jamais serão os mesmos. Em breve, o primeiro participante e o seu treinamento envolverá apenas um experimento prático de uma tarde, em uma aula da graduação, para o seqüenciamento e a comparação de dois genomas de mamíferos.

O genoma humano é fundamentalmente informação, e computadores foram essenciais tanto para a determinação da seqüência quanto para aplicações em biologia e medicina que já estão fluindo a partir destes dados. A computação contribuiu não somente com a sua capacidade natural para processamento e armazenamento de dados, mas também com métodos matematicamente sofisticados, necessários para se obter os resultados. A união entre biologia e ciência da computação criou um novo campo chamado bioinformática.

Hoje, a bioinformática é uma ciência aplicada. Nós utilizamos programas de computador para fazer inferências a partir de dados obtidos da biologia molecular moderna, para fazer conexões entre eles e para derivar predições importantes e relevantes.

Este livro é dirigido a estudantes e cientistas praticantes que precisam saber como ter acesso aos arquivos de dados de genomas e proteínas, às ferramentas que foram desenvolvidas para trabalhar com esses arquivos e aos tipos de questões que esses dados e ferramentas podem responder. Na verdade, existem muitas fontes de informação. Sítios que tratam destes tópicos de bioinformática estão espalhados por toda a *web*. O desafio é selecionar o essencial deste material e descrevê-lo clara e coerentemente, em um nível introdutório.

Assume-se que o leitor já possui algum conhecimento de biologia molecular moderna e alguma facilidade em usar o computador. A proposta deste livro é construir e desenvolver esta base. Ele é adequado como um livro-texto para alunos de graduação avançados ou alunos de pós-graduação iniciantes. Muitos exemplos resolvidos estão integrados ao texto, e referências para importantes sítios na *web* e leituras recomendadas são fornecidas.

Os problemas sugeridos no livro testam e consolidam a compreensão dos conceitos e fornecem oportunidades para praticar habilidades e explorar assuntos adicionais. Três tipos de problemas aparecem no final dos capítulos. Os exercícios são curtos e constituem aplicações diretas do material no texto. As respostas para os exercícios aparecem no final do livro. Os problemas também não precisam de nenhuma informação que não esteja no livro, mas necessitam de respostas mais longas e, em alguns casos, de cálculos mais elaborados. A terceira categoria, *"Weblemas"*, exige acesso à *World Wide Web*. Os *weblemas* são desenvolvidos para fornecer aos leitores prática com as ferramentas necessárias para estudos mais avançados e pesquisa nesta área.

O que tornou possível tentar escrever este livro, neste momento, é a facilidade com que a *World Wide Web* permite acesso tanto aos bancos de dados em si

quanto aos programas que lidam com eles. No passado, era necessário instalar programas e dados em seu próprio computador e executar os cálculos localmente. Claro que isso significava que tudo era dependente dos recursos disponíveis. Agora é possível conduzir todo o trabalho através de uma interface na *web*. Para garantir que os leitores serão capazes de realizar discussões livremente sobre o livro na *web*, descrições e referências a *softwares* comerciais foram evitadas, embora muitos pacotes comerciais sejam de qualidade bastante alta.

Um problema sério com a *web* é a sua inconstância. Sítios vêm e vão, deixando rastros de conexões "mortas" como conseqüência. Existem tantos sítios que é necessário encontrar alguns portais que sejam estáveis – não somente que continuem a existir, mas também que mantenham tanto seu conteúdo quanto suas conexões atualizadas. Eu sugeri alguns desses sítios, mas muitos outros são tão bons quanto esses. O problema não é criar uma longa lista de sítios importantes – isso já foi feito muitas vezes, e é relativamente fácil – mas criar uma lista curta – isso é muito mais difícil!

A programação de computadores é introduzida neste livro baseada na linguagem PERL amplamente disponível. Exemplos de programas simples em PERL aparecem no contexto de problemas biológicos. Muitas tarefas simples em PERL estão associadas a exercícios ou problemas no final dos capítulos.

Estudar PERL muito!

Meu objetivo é que os leitores deste livro desenvolvam:

- Uma compreensão sobre a natureza da grande quantidade disponível de informações detalhadas sobre nós mesmos e sobre outras espécies.
- Uma idéia da abrangência das aplicações da bioinformática nas áreas de biologia molecular, medicina clínica, farmacologia, biotecnologia, ciência forense, antropologia e em outras disciplinas.
- Um conhecimento útil das técnicas por meio das quais, através da *World Wide Web*, obtemos acesso aos dados e aos métodos para analisá-los.
- Uma apreciação do papel dos computadores e da ciência da computação nas investigações e aplicações dos dados.
- Confiança nas suas habilidades básicas para recuperação de informações e cálculos com os dados, e na capacidade de estender essas habilidades para direcionar o seu "trabalho de campo" na *web*.
- Um sentimento de otimismo de que os dados e métodos da bioinformática ocasionarão avanços profundos em nossa compreensão da vida e melhorias na saúde de humanos e de outras formas de vida.

E para onde o leitor deve se dirigir depois? Este livro é projetado como um volume acompanhante – no jargão corrente, um *"prequel"* ou introdução – ao *Introduction to Protein Architecture: The Structural Biology of Proteins*. *Introduction to Protein Science: Architecture, Function, and Genomics* ensaia uma síntese mais geral da área. Claro que esses títulos são apenas recomendações. Outros livros sobre análise de seqüências estendem-se desde aqueles orientados à biologia até os da área de ciência da computação. O objetivo é que cada leitor venha a reconhecer seus próprios interesses e esteja equipado para levá-los adiante.

A.M.L.

Sumário

1 Introdução *21*

A vida no espaço e no tempo *23*

Evolução é a mudança no decorrer do tempo no mundo dos organismos vivos *24*

Dogmas: central e periférico *26*

Observáveis e arquivos de dados *29*

O fluxo da informação na bioinformática *32*

Organização, anotação e controle de qualidade *33*

A World Wide Web *34*

Publicação eletrônica *35*

Computadores e a ciência da computação *36*

Programação *37*

Classificação e nomenclatura biológica *41*

O uso de seqüências na determinação de relações filogenéticas *44*

O uso de SINES e LINES na derivação de relações filogenéticas *50*

Pesquisa por seqüências similares em bancos de dados: PSI-BLAST *52*

Introdução à estrutura de proteínas *60*

A natureza hierárquica da arquitetura de proteínas *61*

Classificação de estruturas de proteínas *64*

Predição e engenharia da estrutura de proteínas *71*

Critical Assessment of Structure Prediction (CASP) *72*

Engenharia de proteínas *72*

Proteômica *72*

Microarranjos de DNA *73*

Espectrometria de massa *74*

Biologia de sistemas *74*

Implicações clínicas *75*

O futuro *77*

Leituras recomendadas *77*

Exercícios, problemas e weblemas *79*

2 Organização e evolução de genomas *87*

Genomas e proteomas *88*

Genes *89*

Proteomas *91*

Espionando a transmissão da informação genética *92*

Mapeamentos entre os mapas *97*

Mapas de alta resolução *98*

Selecionando genes em genomas *100*
Genomas de procariotos *101*
O genoma da bactéria *Escherichia coli* *102*
O genoma da arqueobactéria *Methanococcus jannaschii* *105*
O genoma de um dos organismos mais simples: *Mycoplasma genitalium* *106*
Genomas de eucariotos *107*
O genoma de *Saccharomyces cerevisiae* (fermento de padeiro) *109*
O genoma de *Caenorhabditis elegans* *112*
O genoma de *Drosophila melanogaster* *114*
O genoma de *Arabidopsis thaliana* *115*
O genoma do *Homo sapiens* (o genoma humano) *116*
Genes que codificam proteínas *117*
Seqüências de repetições *119*
RNA *120*
Polimorfismos de um único nucleotídeo (SNPs) *121*
Diversidade genética na antropologia *122*
Diversidade genética e identificação pessoal *123*
Análise genética da domesticação do gado *124*
Evolução de genomas *124*
Por gentileza, passem os genes: transmissão horizontal de genes *128*
Genômica comparativa de eucariotos *129*
Leituras recomendadas *131*
Exercícios, problemas e weblemas *132*

3 Bancos de dados e recuperação de informação *137*

Introdução *138*
Indexação de bancos de dados e especificação de termos de pesquisa *138*
Monitoração dos resultados da pesquisa *140*
Análise dos dados recuperados *141*
Os bancos de dados *141*
Bancos de dados de seqüências de ácidos nucléicos *142*
Bancos de dados de genomas *144*
Bancos de dados de seqüências de proteínas *144*
Bancos de dados de estruturas *148*
Bancos de dados especializados ou "boutique" *155*
Bancos de dados de padrões de expressão e de proteômica *156*
Bancos de dados de vias metabólicas *158*
Bancos de dados bibliográficos *159*
Avaliações sobre bancos de dados e servidores de biologia molecular *159*
Acessos a bancos de dados *160*
Acesso a bancos de dados de biologia molecular *161*
ENTREZ *161*
O Sequence Retrieval System (SRS) *168*
O Protein Information Resource (PIR) *169*

EXPASy – Expert Protein Analysis System 170
Ensembl 171
Para onde vamos a partir daqui? *172*
Leituras recomendadas *172*
Exercícios, problemas e weblemas *173*

4 Alinhamentos e árvores filogenéticas *177*

Introdução ao alinhamento de seqüências *178*
O gráfico de pontos *180*
Gráficos de pontos e alinhamentos de seqüências *185*
Medidas de similaridade de seqüências *191*
Esquemas de pontuação *191*
Calculando o alinhamento de duas seqüências *195*
Variações e generalizações *195*
Métodos aproximados para a análise rápida de bancos de dados *196*
O algoritmo de programação dinâmica para o alinhamento ótimo de seqüências par a par *196*
A significância de alinhamentos *202*
Alinhamento múltiplo de seqüências *206*
Aplicações de alinhamentos múltiplos de seqüências em consultas a banco de dados *208*
Perfis *209*
PSI-BLAST *211*
Modelos ocultos de Markov *213*
Filogenia *218*
Árvores filogenéticas *223*
Métodos de agrupamento *225*
Métodos cladísticos *226*
O problema da variação das taxas de evolução *227*
Considerações computacionais *228*
Leituras recomendadas *229*
Exercícios, problemas e weblemas *230*

5 Estrutura de proteínas e descoberta de fármacos *239*

Introdução *240*
Estabilidade e enovelamento de proteínas *243*
O mapa de Sasisekharan-Ramakrishnan-Ramachandran descreve as conformações permitidas da cadeia principal *243*
As cadeias laterais *245*
Estabilidade e desnaturação de proteínas *245*
Enovelamento de proteínas *248*
Aplicações da hidrofobicidade *249*
Sobreposição de estruturas e alinhamentos estruturais *253*
DALI (Distance-matrix ALIgnment) *255*
Evolução de estruturas de proteínas *256*

Classificação de estruturas de proteínas *258*
SCOP *259*
Predição e modelagem de estrutura de proteínas *260*
Critical Assessment of Structure Prediction (CASP) *262*
Predição de estrutura secundária *264*
Modelagem por homologia *270*
Reconhecimento de padrão de enovelamento ou topologia *272*
Cálculos de energia conformacional e dinâmica molecular *275*
ROSETTA *279*
LINUS *279*
Atribuição de estruturas de proteínas a genomas *282*
Predição de função de proteínas *285*
Divergência de função: ortólogas e parálogas *286*
Descoberta e desenvolvimento de fármacos *289*
O composto líder *291*
A bioinformática na descoberta e no desenvolvimento de fármacos *293*
Leituras recomendadas *304*
Exercícios, problemas e weblemas *305*

6 Proteômica e biologia de sistemas *311*

Microarranjos de DNA *313*
Análise dos dados de microarranjos *315*
Espectrometria de massa *321*
Identificação dos componentes de uma mistura complexa *321*
Seqüenciamento de proteínas por espectrometria de massa *324*
Análise de seqüências de genomas por espectrometria de massa *326*
Biologia de sistemas *331*
Redes e grafos *333*
Estrutura e dinâmica de redes *338*
Complexos e agregados de proteínas *340*
Propriedades de complexos proteína-proteína *341*
Redes de interações protéicas *344*
Redes reguladoras *349*
Estruturas de redes reguladoras *350*
Biologia estrutural de redes reguladoras *356*
Leituras recomendadas *359*
Exercícios, problemas e weblemas *359*
Conclusões *365*
Respostas aos exercícios *367*
Glossário *373*
Índice *377*
Gravuras coloridas *entre as páginas 192 e 193*

Plano do livro

- O Capítulo 1 prepara o terreno ao introduzir todos os conceitos principais que serão utilizados no livro: seqüência e estrutura de DNA e proteínas, genomas e proteomas, bancos de dados e recuperação de informações, a *World Wide Web* e programação de computadores. Antes de desenvolver os tópicos individuais em detalhes, é importante perceber a estrutura geral de suas interações.
- O Capítulo 2 apresenta a natureza dos genomas individuais, incluindo o genoma humano, e as suas relações do ponto de vista biológico.
- O Capítulo 3 trabalha com as habilidades básicas necessárias para se utilizar a *web* na bioinformática. Ele descreve os arquivos de bancos de dados e orienta o leitor através de algumas sessões práticas envolvendo a recuperação de informações de alguns dos principais bancos de dados de biologia molecular.
- O Capítulo 4 trabalha a análise de relações entre seqüências – alinhamentos e árvores filogenéticas. Estes métodos formam a base de alguns dos principais desafios computacionais da bioinformática: detectar parentescos distantes, compreender relações entre genomas de diferentes organismos e investigar o curso da evolução das espécies no nível molecular.
- O Capítulo 5 adentra as três dimensões espaciais, tratando os problemas de estrutura e enovelamento de proteínas. Seqüência e estrutura de proteínas devem ser vistos como cúmplices, com a bioinformática desenvolvendo métodos para que se possa ir da seqüência à estrutura, e vice-versa, da maneira mais natural possível. A compreensão de estruturas de proteínas em detalhes é essencial para a determinação de seus mecanismos de ação e para aplicações clínicas e farmacológicas.
- O Capítulo 6 discute proteômica e biologia de sistemas, incluindo os novos métodos rápidos e de produção em larga escala de dados sobre a expressão e distribuição de proteínas nas células, bem como as tentativas de se sintetizar essas informações com o objetivo de revelar padrões de organização.

CAPÍTULO 1

Introdução

Sumário do capítulo

A vida no espaço e no tempo 23

Evolução é a mudança no decorrer do tempo no mundo dos organismos vivos 24

Dogmas: central e periférico 26

Observáveis e arquivos de dados 29

O fluxo da informação na bioinformática 32

Organização, anotação e controle de qualidade 33

A *World Wide Web* 34

Publicação eletrônica 35

Computadores e a ciência da computação 36

Programação 37

Classificação e nomenclatura biológica 41

O uso de seqüências na determinação de relações filogenéticas 44

O uso de SINES e LINES na derivação de relações filogenéticas 50

Pesquisa por seqüências similares em bancos de dados: PSI-BLAST 52

Introdução à estrutura de proteínas 60

A natureza hierárquica da arquitetura de proteínas 61

Classificação de estruturas de proteínas 64

Predição e engenharia da estrutura de proteínas 71

Critical Assessment of Structure Prediction (CASP) 72

Engenharia de proteínas 72

Proteômica 72

Microarranjos de DNA 73

Espectrometria de massa 74

Biologia de sistemas 74

Implicações clínicas 75

O futuro 77

Leituras recomendadas 77

Exercícios, problemas e weblemas 79

A biologia tem sido tradicionalmente uma ciência mais de observação do que de dedução. Apesar de os recentes desenvolvimentos não terem alterado esta premissa básica, a natureza da informação foi modificada radicalmente. Pode-se argumentar que, até há pouco tempo, todas as observações biológicas eram fundamentalmente subjetivas – reconhecidamente com diferentes graus de precisão, alguns inclusive bastante elevados. Entretanto, nos últimos anos, os dados se tornaram não apenas muito mais quantitativos e precisos como, no caso de seqüências de nucleotídeos e aminoácidos, se tornaram *discretos*. É possível determinar a seqüência genômica de um único organismo ou clone não apenas de forma completa, mas também, em princípio, de forma *exata*. Erros experimentais não podem ser inteiramente evitados, mas, no seqüenciamento moderno de genomas, eles são extremamente baixos.

Não que isso tenha convertido a biologia em uma ciência dedutiva. A vida segue os princípios da física e química, mas a vida ainda é muito complexa e muito dependente de contingentes históricos para que suas propriedades possam ser deduzidas em detalhes a partir de princípios básicos. Se tal característica empobrece ou enriquece a biologia, é uma questão de opinião.

Uma segunda propriedade óbvia dos dados de bioinformática é o seu grande volume. Atualmente, os bancos de dados de seqüências de nucleotídeos contêm 80.000×10^6 bases, ou, abreviadamente, 80.000 Mpb.* Se utilizarmos o tamanho aproximado do genoma humano – 3×10^9 letras – como unidade, esses bancos de dados conteriam 26 equivalentes ao genoma humano (*Human Genome Equivalents*, ou 26 ***huges***, um nome apropriado). Para um padrão de comparação mais abrangente, 1 *huge* equivale ao número total de caracteres impressos em seis anos completos de edições do jornal *The New York Times*. O banco de dados de estruturas de macromoléculas contém 30.000 entradas, ou seja, as coordenadas tridimensionais completas de proteínas, com tamanho médio de aproximadamente 400 resíduos.** Os diferentes bancos de dados não são apenas extensos, mas os seus tamanhos crescem a taxas bastante elevadas. A Figura 1.1 mostra o crescimento ao longo da última década do GenBank (banco de dados de seqüências de ácidos nucléicos) e do Protein Data Bank (banco de dados de estruturas de macromoléculas). Como se pode notar, extrapolações podem ser imprecisas.

Esta qualidade e quantidade de dados encorajaram os cientistas a traçarem objetivos consideravelmente ambiciosos:

- Afirmar que "viram a vida clara e completamente". Ou seja, o entendimento de aspectos integrados da biologia dos organismos, vistos como sistemas complexos coerentes.
- Inter-relacionar seqüência, estrutura tridimensional, padrões de expressão, interações e função de proteínas individuais, ácidos nucléicos e complexos proteína-ácidos nucléicos.
- Integrar os dados sobre diferentes aspectos da vida de uma célula ou de um organismo numa descrição, em termos de "sistema", da sua estrutura e dinâmica.
- Usar dados acerca de organismos contemporâneos como base para inferências sobre o passado e o futuro – no passado, para deduzir eventos da

* N. de T. Em abril de 2007, este número já ultrapassava a casa dos 100 bilhões ou 10^9 letras ou bases!

** N. de T. Em abril de 2007, este número já ultrapassava a casa das 39.000 entradas apenas para proteínas!

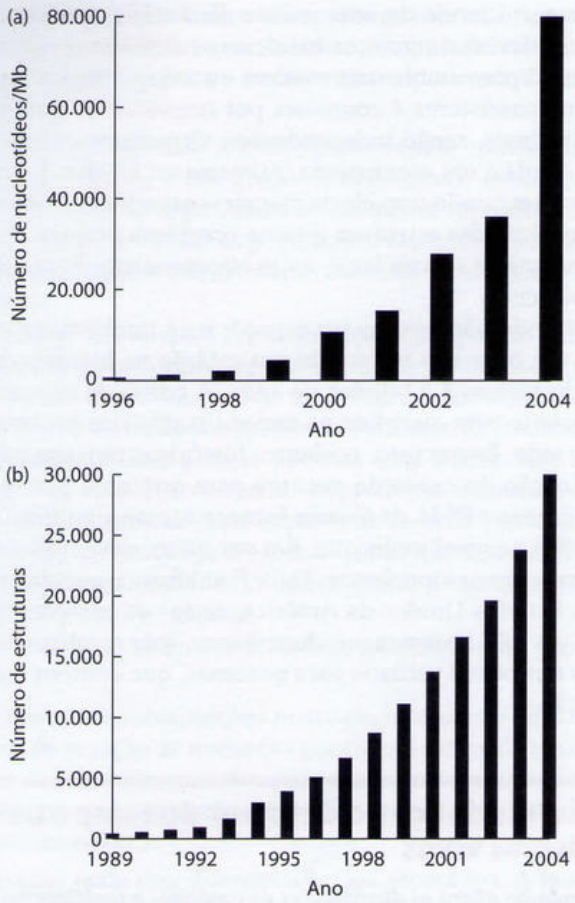


Figura 1.1 (a) Crescimento do GenBank, um banco de dados de arquivos de seqüências genéticas do US National Center for Biotechnology Information (NCBI). (b) Crescimento do Protein Data Bank, um arquivo de estruturas tridimensionais de macromoléculas biológicas.

história evolucionária e, no futuro, para nortear modificações científicas de sistemas biológicos.

- Apoiar aplicações nas áreas de medicina, agricultura e tecnologia.

A vida no espaço e no tempo

Definição VIDA

É difícil definir "vida", e pode ser necessário modificar sua definição – ou viver desconfortavelmente com a atual – à medida que os computadores evoluem em capacidade de processamento e a interface entre *in vivo* e *in silico* se torna mais tênue. Por hora, experimente esta definição: um organismo biológico é um dispositivo de ocorrência natural, que se auto-reproduz e é capaz de manipular, de forma controlada, matéria, energia e informação.

De uma perspectiva mais ampla, a vida na Terra é um sistema auto-replicativo complexo, distribuído no tempo e no espaço. É da maior importância que ela seja composta, em grande parte, por organismos *distintos*, cada qual com tempo de vida finito e, na maioria dos casos, com características singulares.

Espacialmente, iniciando de uma grande distância e aproximando progressivamente, é possível distinguir, na biosfera, os **ecossistemas** locais, estáveis até que suas condições ambientais mudem ou até que sejam invadidos. Cada espécie em um ecossistema é composta por **organismos** que desempenham atividades individuais, senão independentes. Organismos são compostos por **células**. Cada célula é um ecossistema intimamente localizado, não isolado do ambiente, mas interagindo com ele de maneiras específica e controlada. Células eucarióticas contêm uma estrutura interna complexa própria, incluindo o núcleo e outras organelas subcelulares, e um citoesqueleto. E, por fim, chegamos ao nível das moléculas.

A vida se estende não apenas no espaço, mas também no tempo. O que vemos hoje é um pequeno retrato de um estágio na história da vida que se estende há pelo menos 3,5 bilhões de anos. A teoria da seleção natural tem sido extremamente bem-sucedida na racionalização dos processos de desenvolvimento da vida. Entretanto, acidentes históricos têm um papel dominante na determinação do curso de eventos para que uma predição detalhada seja possível. E nem o DNA de fósseis fornece acesso significativo a qualquer registro histórico no nível molecular. Em vez disso, devemos tentar ler o passado em genomas contemporâneos. Felix Frankfurter, juiz da Suprema Corte de Justiça dos Estados Unidos da América, certa vez escreveu que "a constituição americana não é apenas um documento, mas também um registro da história". Isso também é verdade para genomas, que contêm registros da sua própria evolução.

Evolução é a mudança no decorrer do tempo no mundo dos organismos vivos

O processo de evolução altera as distribuições de genótipos e fenótipos em gerações sucessivas. O **genótipo** corresponde às informações genéticas de um organismo, a seqüência do seu genoma. Todas as características que podem ser observadas em um organismo – de forma macroscópica e bioquímica – compreendem o **fenótipo**. O genótipo é herdado de um dos pais, ou de ambos, e está sujeito a modificações por mutações ou por transferência lateral de material genético. O fenótipo depende do genótipo, que controla o desenvolvimento do organismo sob a influência do seu ambiente.

A assimetria entre o genótipo e o fenótipo é o motor da evolução:

- As alterações no genótipo são hereditárias. Efeitos do ambiente ou do estilo de vida sobre o fenótipo – por exemplo, uma melhor nutrição levando a um aumento da massa corporal, ou os efeitos debilitantes de uma doença ou lesões – não são diretamente herdáveis.
- Durante o desenvolvimento de qualquer organismo, o genótipo limita o fenótipo. O fenótipo não influencia o genótipo.
- Muitos genótipos podem criar o mesmo fenótipo:
  - Muitas mutações nos genes codificadores de proteínas mantêm a seqüência de aminoácidos inalterada, ou acarretam modificações sem efeito na função.
  - **Alelos** são diferentes formas (seqüências) de um mesmo gene. Qualquer organismo que contenha duas ou mais cópias de um gene pode repetir o mesmo alelo (homozigoto) ou conter diferentes alelos (heterozigoto). Ho-

Em mamíferos, ~20% dos *loci* são heterozigotos. ?

mozigotos e heterozigotos têm diferentes genótipos, mas, se um alelo for dominante, e se um único gene apresentar controle exclusivo sobre uma característica, homozigotos e heterozigotos podem apresentar o mesmo fenótipo.

Em quais níveis a evolução atua? A maior parte da vida consiste em organismos distintos. Uma **população** é um grupo de organismos semelhantes que interagem: uma população de organismos sexualmente reprodutivos pode cruzar entre si; os indivíduos, em todas as populações, competem por recursos. A evolução altera a composição e a distribuição do conjunto de genes e fenótipos em uma população.

Qual é o mecanismo de evolução? Em uma população, podem surgir indivíduos com uma variedade de genótipos, apresentando uma variedade correspondente de fenótipos. Apesar de a evolução não ter influência *direta* sobre o genótipo, indivíduos com diferentes fenótipos apresentam sucesso diferenciado na reprodução. Como resultado, a nova geração pode ter uma distribuição alterada de genótipos e fenótipos. A **seleção natural** – reprodução acentuada dos indivíduos "mais aptos" – é o mecanismo mais importante de evolução. Outro mecanismo é a **deriva gênica**, ou mudança aleatória nas freqüências alélicas, não como resultado de um processo seletivo. A deriva gênica é especialmente importante em populações pequenas e isoladas.

Os mecanismos que produzem variedade genética criam um *potencial* para evolução:

- **Mutações**, tais como substituições pontuais, inserções e deleções, e transposições. A taxa de geração de mutações pontuais é estimada em cerca de 10^{-12} a 10^{-10} por par de bases por geração. (Isso *não* é o mesmo que a taxa de substituição alélica em uma população. Mutações apenas propõem candidatos para mudança evolucionária.)
- A **recombinação** pode unir diferentes *loci* ou separá-los. A recombinação em um gene pode criar um novo alelo. A recombinação entre genes pode alterar a relação entre genes e elementos regulatórios.
- **Duplicação gênica**, seguida por divergência.
- **Fluxo gênico**, a partir da mistura de populações, ou da transferência lateral de genes entre espécies.

A evolução pode aumentar ou diminuir a variedade do reservatório de genes. Se uma mutação distinta confere vantagens seletivas apenas no estado homozigoto, o gene pode se espalhar em toda a população. A adoção do alelo por todos os membros da população pode *diminuir* a variedade no reservatório genético. Se um gene surge e confere vantagens seletivas apenas no estado heterozigoto, o conjunto pode passar a apresentar maior variedade. Algumas mutações originam alelos recessivos que são deletérios apenas quando em estado homozigoto. Esses alelos são difíceis de remover de uma população, especialmente se os heterozigotos apresentarem alguma vantagem compensatória, como no caso da anemia falciforme, na qual indivíduos heterozigotos apresentam maior resistência à malária.

Microevolução é um termo que se refere a mudanças relativamente pequenas em uns poucos genes, levando, na maioria dos casos, a alterações relativamente pequenas nos fenótipos. A microevolução afeta indivíduos *em uma* população. Técnicas modernas nos permitem rastrear a microevolução em nível molecular, por meio da examinação de seqüências genômicas e padrões

de expressão de proteínas. **Macroevolução** se refere a mudanças em larga escala, que ocorrem em uma população como um todo, incluindo a formação de novas espécies. O registro fóssil fornece uma história (parcial) da macroevolução, com o uso de métodos geológicos para a datação de eventos. A anatomia e a fisiologia comparativas, bem como a embriologia, fornecem informações adicionais.

As observações da micro e da macroevolução se complementam. As seqüências genômicas auxiliam na classificação de espécies. O registro fóssil permite a datação de eventos passados que tiveram conseqüências na informação molecular que observamos atualmente. Um grande desafio para a biologia moderna é o entendimento de como eventos em larga escala, tal como o desenvolvimento de novas espécies, podem ocorrer como resultado de eventos microevolucionários.

Dogmas: central e periférico

O arquivo de informações em cada organismo – o plano de desenvolvimento e atividades em potencial – é o material genético, DNA, ou, em alguns vírus, o RNA. Moléculas de DNA são cadeias longas, *lineares*, contendo uma mensagem em um alfabeto de quatro letras (ver Quadro). Mesmo para microrganismos a mensagem é longa, tipicamente com 10^6 caracteres. Implícitos na estrutura do DNA estão os mecanismos para auto-replicação e tradução dos genes em proteínas. A dupla hélice, e sua autocomplementaridade interna, proporcionando uma replicação correta, são bem conhecidas (ver Gravura I). A replicação quase perfeita é essencial para a estabilidade da herdabilidade, mas algumas imperfeições neste processo de replicação, ou mecanismos de importação de material genético não-próprio, são também necessários, caso contrário a evolução não poderia ocorrer em organismos assexuados.

As fitas na dupla hélice de DNA são antiparalelas; as direções ao longo de cada uma das fitas são indicadas pelas extremidades 3' e 5' (pelas posições no anel desoxirribose). Na tradução para proteínas, a seqüência de DNA é sempre lida na direção 5' → 3'.

A implementação da informação genética ocorre inicialmente com a síntese de RNA e proteínas. As proteínas são as moléculas responsáveis pela maior parte da estrutura e atividade dos organismos. Nossos cabelos, músculos, enzimas digestivas e anticorpos são todos proteínas. Tanto os ácidos nucléicos como as proteínas são moléculas que se apresentam como cadeias longas e lineares. O "código" genético é de fato uma codificação: tripletos de letras sucessivas da seqüência de DNA especificam aminoácidos consecutivos; porções da seqüência de DNA codificam seqüências de aminoácidos de proteínas. Tipicamente, proteínas são compostas de 200 a 400 aminoácidos, o que exige de 600 a 1.200 letras de mensagens de DNA expresso para especificá-las. A síntese de moléculas de RNA, como, por exemplo, os componentes do RNA do ribossomo, é também determinada por seqüências de DNA.

Sydney Brenner estabeleceu a distinção entre "supérfluo" e "lixo": o lixo é descartável, enquanto o supérfluo é mantido.

No entanto, nem todo o DNA é expresso como proteínas ou RNA estrutural. Muitos dos genes nos organismos superiores contêm seqüências internas não traduzidas, ou íntrons. Algumas regiões da seqüência de DNA atuam como mecanismos de controle, e uma porção substancial do genoma dos organismos superiores aparenta ser "supérflua". (O que pode significar simplesmente que ainda não compreendemos sua função.)

Os quatro nucleotídeos de ocorrência natural no DNA (RNA)

a adenina g guanina c citosina t timina (u uracila)

Os vinte aminoácidos de ocorrência natural em proteínas

Aminoácidos não-polares

G glicina	A alanina	P prolina	V valina
I isoleucina	L leucina	F fenilalanina	M metionina

Aminoácidos polares

S serina	C cisteína	T treonina	N asparagina
Q glutamina	H histidina	Y tirosina	W triptofano

Aminoácidos carregados

D ácido aspártico	E ácido glutâmico	K lisina	R arginina

Outras classificações de aminoácidos também podem ser úteis. Por exemplo, histidina, fenilalanina, tirosina e triptofano são aromáticos e desempenham papéis estruturais especiais em proteínas de membrana.

Os nomes dos aminoácidos são freqüentemente abreviados com suas três primeiras letras, como Gli para glicina, exceto para isoleucina, asparagina, glutamina e triptofano, que são abreviados como Ile, Asn, Gln e Trp, respectivamente. O aminoácido raro selenocisteína tem, como abreviação de três letras, Sec e código de uma letra U.

Convencionou-se escrever os nucleotídeos em letras minúsculas e os aminoácidos em letras maiúsculas. Assim, atg = adenina-timina-guanina, e ATG = alanina-treonina-glicina.

No DNA, as moléculas que compõem o alfabeto são quimicamente similares, e a estrutura do DNA é, em uma primeira aproximação, uniforme (embora algumas interações DNA-proteína causem distorções na estrutura do DNA). Proteínas e RNAs estruturais, ao contrário, apresentam ampla variedade de conformações tridimensionais. Essas conformações são necessárias para garantir o desempenho de seus diversos papéis funcionais.

A seqüência de aminoácidos de uma proteína determina sua estrutura tridimensional. Para cada seqüência de aminoácido natural, há um único *estado nativo* estável, o qual, sob condições adequadas, é adotado espontaneamente. Se uma proteína purificada é aquecida, ou submetida a condições diferentes do meio fisiológico normal, ela irá se "desenovelar" em uma estrutura desordenada e biologicamente inativa. (É por essa razão que nossos corpos têm mecanismos para a manutenção quase constante de nossas condições internas.) Quando as condições normais são restauradas, as moléculas protéicas geralmente reassumem a sua estrutura nativa, indistinguível do estado original.

As funções das proteínas dependem de elas adotarem a estrutura tridimensional do seu estado nativo. Por exemplo, a estrutura nativa de uma enzima pode apresentar uma cavidade na sua superfície, que se liga a uma pequena molécula

Códigos genéticos alternativos são observados em organelas – cloroplastos e mitocôndrias – e em algumas espécies.

O código genético padrão

ttt	Phe	tct	Ser	tat	Tir	tgt	Cis
ttc	Phe	tcc	Ser	tac	Tir	tgc	Cis
tta	Leu	tca	Ser	taa	parada	tga	parada
ttg	Leu	tcg	Ser	tag	parada	tgg	Trp
ctt	Leu	cct	Pro	cat	His	cgt	Arg
ctc	Leu	ccc	Pro	cac	His	cgc	Arg
cta	Leu	cca	Pro	caa	Gln	cga	Arg
ctg	Leu	ccg	Pro	cag	Gln	cgg	Arg
att	Ile	act	Tre	aat	Asn	agt	Ser
atc	Ile	acc	Tre	aac	Asn	agc	Ser
ata	Ile	aca	Tre	aaa	Lis	aga	Arg
atg	Met	acg	Tre	aag	Lis	agg	Arg
gtt	Val	gct	Ala	gat	Asp	ggt	Gli
gtc	Val	gcc	Ala	gac	Asp	ggc	Gli
gta	Val	gca	Ala	gaa	Glu	gga	Gli
gtg	Val	gcg	Ala	gag	Glu	ggg	Gli

De uma para três dimensões

O enovelamento espontâneo de proteínas para formar seus estados nativos é o ponto em que a natureza realiza o grande salto do mundo unidimensional dos genes e seqüências de proteínas para o mundo tridimensional em que vivemos. Existe um paradoxo: a tradução das seqüências de DNA em seqüências de aminoácidos é muito simples de descrever de maneira lógica; ela é especificada pelo código genético. O enovelamento de uma cadeia polipeptídica em uma estrutura tridimensional precisa é muito difícil de explicar de maneira lógica. Entretanto, a tradução exige a maquinaria imensamente complicada dos ribossomos, dos tRNAs e das moléculas associadas; todavia, o enovelamento de proteínas ocorre espontaneamente.

e a coloca próximo de seus resíduos catalíticos. Muitos mecanismos reguladores dependem da ligação de proteínas a outras proteínas ou ao DNA. Assim, temos um paradigma:

- A seqüência de DNA determina a seqüência da proteína
- A seqüência da proteína determina a estrutura da proteína
- A estrutura da proteína determina a função da proteína
- Mecanismos reguladores, incluindo mas não limitado ao controle de padrões de expressão, fornecem as quantidades corretas das funções corretas, nos momentos e nos locais corretos

Grande parte da atividade organizada da bioinformática está focalizada na análise de dados relacionados a esses processos.

Até o presente momento, esse paradigma não inclui níveis maiores do que o da estrutura e organização molecular, incluindo, por exemplo, questões sobre como tecidos se tornam especializados no curso do desenvolvimento ou, de forma mais geral, como efeitos ambientais exercem controle sobre eventos genéticos. Em alguns casos de ciclos de retroalimentação simples, já se compreende, em nível molecular, como um aumento na quantidade de um reagente acarreta um aumento na produção de uma enzima que catalisa sua transformação. Os programas de desenvolvimento durante a vida de um organismo são mais complexos. Estes problemas fascinantes sobre o fluxo e o controle da informação em um organismo agora são acessíveis ao estado da arte da bioinformática. O tópico de **biologia de sistemas** está centrado na integração e no controle das atividades de células e organismos.

Observáveis e arquivos de dados

Um banco de dados inclui um arquivo de informações, uma organização lógica ou "estruturada" dessas informações e ferramentas para se ter acesso a elas. Os bancos de dados da biologia molecular contêm seqüências de ácidos nucléicos e de proteínas, estruturas e funções de macromoléculas, padrões de expressão, redes de vias metabólicas e cascatas de regulação. Eles incluem:

- Bancos de dados primários de arquivos de informações biológicas:
  - Seqüências de DNA e proteínas, incluindo anotação
  - Variações, tais como compilações de haplótipos
  - Estruturas de ácidos nucléicos e proteínas, incluindo anotação
  - Bancos de dados específicos para organismos, incluindo bancos de dados de genomas
  - Bancos de dados de padrões de expressão protéica
  - Bancos de dados de rotas metabólicas
  - Bancos de dados de padrões de interação e de vias reguladoras
- Bancos de dados derivados ou secundários: estes contêm informações obtidas dos bancos de dados primários e das análises dos seus conteúdos. Por exemplo:
  - motivos de seqüências protéicas ("padrões de assinatura" característicos de famílias de proteínas)
  - mutações e variantes nas seqüências de DNA e de proteínas
  - classificações ou relações (conexões e características comuns das entradas dos arquivos; por exemplo, um banco de dados de conjuntos de famílias de seqüências de proteínas, ou uma classificação hierárquica de padrões de enovelamento de proteínas)
- Bancos de dados bibliográficos
- Bancos de dados de sítios na *web*:
  - bancos de bancos de dados contendo informações biológicas
  - conexões entre bancos de dados

Recursos *na web*: Seqüências de ácidos nucléicos e de proteínas

O arquivo de seqüências de ácidos nucléicos é mantido por uma parceria entre três entidades: o *GenBank*, situado no US National Center for Biotechnology Information (NCBI), em Bethesda, Maryland, Estados Unidos; o *EMBL Nucleotide Sequence Database*, localizado no European Bioinformatics Institute (EBI), em Hinxton, no Reino Unido; e o *The Center for Information Biology e DNA DataBank of Japan*, no National Institute of Genetics em Mishima, Japão. Os três sítios trocam informações sobre novas submissões de dados diariamente, para garantir que todos tenham o mesmo conteúdo. Entretanto, o formato, a anotação e as conexões inerentes diferem entre as entradas correspondentes fornecidas pelos diferentes bancos de dados.

O arquivo de seqüências de aminoácidos de proteínas, ora determinado quase exclusivamente pela tradução de seqüências de genes, é mantido pelo *United Protein Database* (*UniProt*), uma junção dos bancos de dados do *SWISS-PROT*, do *The Protein Identification Resource (PIR)* e do *Translated EMBL (TrEMBL)*.

Associadas a estes arquivos existem ferramentas para a seleção e recuperação de seqüências. O *Sequence Retrieval System (SRS)*, um produto da Lion Bioscience AG, está disponível gratuitamente para uso acadêmico por meio do European Bioinformatics Institute e de inúmeros outros sítios-espelho. O NCBI, nos Estados Unidos, oferece o *ENTREZ*. Ambos permitem pesquisas paralelas em múltiplos arquivos de dados.

Muitos projetos de seqüenciamento de genomas completos mantêm bancos de dados focalizados em espécies individuais. Exemplos notáveis são o *ENSEMBL* (Sanger Centre, Hinxton, Reino Unido) e os navegadores da Universidade da Califórnia, em Santa Cruz, Estados Unidos, para o genoma humano e outras espécies.

Muitos bancos de dados secundários agrupam famílias de proteínas ou subunidades com base na similaridade entre suas seqüências. Um banco de dados "guarda-chuva", o Interpro, integra os conteúdos, as características e a anotação de diversos bancos de dados individuais de famílias de proteínas, domínios e sítios funcionais; além disso, contém conexões para outros bancos, incluindo a classificação funcional do Gene Ontology Consortium™. O Interpro pretende incorporar bancos de dados adicionais. ("Resistir é inútil.")

O mecanismo de acesso a um banco de dados é o conjunto de ferramentas para responder questões, como:

- "O banco de dados contém as informações de que eu preciso?" (Exemplo: Em quais bancos de dados posso encontrar as seqüências de aminoácidos das proteínas álcool desidrogenases?)
- "Como posso organizar as informações selecionadas de banco de dados de maneira útil?" (Exemplo: Como posso compilar uma lista de seqüências de globinas, ou melhor, uma tabela com seqüências de globinas alinhadas?)
- Índices de bancos de dados são úteis em questões como "Onde posso encontrar alguma informação específica?" (Exemplo: Quais bancos de dados contêm a seqüência de aminoácidos da proteína tripsina de porco-espinho?) É claro que, se eu souber e puder especificar exatamente o que quero, o problema será relativamente simples.

Um banco de dados sem meios efetivos de acesso é apenas um cemitério de informações. Como ter acesso efetivo é uma das questões do desenvolvimento de bancos de dados que, idealmente, deve permanecer oculta aos usuários. Tornou-se claro que um acesso eficaz não pode ser fornecido apenas ao se disponibilizar um sistema de consulta sobre arquivos desordenados. Ao contrário, a organização lógica do armazenamento da informação deve ser elaborada já com os meios de acesso em mente – que tipos de questões os usuários irão formular – e a estrutura do arquivo deve se moldar ao *software* utilizado para recuperar a informação.

Uma variedade de consultas a bancos de dados pode surgir na bioinformática. Estas incluem:

(1) Dada uma seqüência, ou fragmento de uma seqüência, encontrar seqüências no banco de dados que sejam similares à seqüência ou fragmento. Este é um problema fundamental na bioinformática. Compartilhamos esses problemas de pareamento de seqüências (conjunto consecutivo de caracteres) com muitos campos da ciência da computação. Por exemplo, programas de edição e processamento de texto possuem funções de procura de seqüências de caracteres.

(2) Dada a estrutura de uma proteína, ou parte de uma estrutura protéica, encontrar estruturas de proteínas no banco de dados que sejam similares à estrutura ou parte dela. Esta é a generalização do problema de pareamento de seqüências em três dimensões.

(3) Dada a seqüência de uma proteína de estrutura desconhecida, encontrar *estruturas* no banco de dados que adotem estruturas tridimensionais (3D) similares. A procura nos bancos de dados de seqüências por proteínas com seqüências similares à seqüência fornecida pode ser tentadora: pois, se duas proteínas apresentarem seqüências suficientemente similares, elas terão estruturas similares. Entretanto, o contrário não é verdadeiro, e pode-se ter esperanças de desenvolver ferramentas de pesquisa mais poderosas que encontrem proteínas de estruturas similares mesmo que suas seqüências tenham divergido além do ponto de poderem ser reconhecidas como similares com base apenas na comparação de suas seqüências.

(4) Dada a estrutura de uma proteína, encontrar *seqüências* no banco de dados que correspondam a estruturas similares. Novamente, pode-se ficar tentado a usar a estrutura para consultar o banco de dados de estruturas, porém isso somente resultará em sucesso limitado, pois existem muito mais seqüências conhecidas do que estruturas. É necessário, então, um método que possa selecionar estruturas a partir de seqüências.

Os problemas (1) e (2) já estão resolvidos; pesquisas como estas são realizadas milhares de vezes por dia. Já os problemas (3) e (4) são campos ativos de pesquisa.

Tarefas de maior complexidade surgem quando se deseja estudar relações entre informações contidas em diferentes bancos de dados. Isso exige conexões que facilitem o acesso simultâneo a diversos bancos de dados. Aqui está um exemplo: "Para quais proteínas de estrutura conhecida, envolvidas em doenças da biossíntese de purinas em humanos, existem proteínas relacionadas em fungos?" Estamos determinando condições sobre estrutura conhecida, função especificada, detecção de relação, correlação com doenças e espécies determinadas. Atualmente, a qualidade de um banco de dados depende não apenas da informação que contém, mas também da efetividade de suas conexões com outras fontes de informação. A importância crescente do acesso simultâneo aos bancos de dados levou à pesquisa

sobre a integração entre eles – como os bancos de dados podem "conversar entre si" sem sacrificar suas liberdades de estruturar seus próprios dados e de maneira apropriada às características individuais das informações que contêm.

Um problema que ainda não surgiu na biologia molecular é o controle de atualizações dos arquivos. O banco de dados de reservas de uma companhia aérea deve evitar que diferentes agentes vendam o mesmo assento para diferentes passageiros. Na bioinformática, os usuários podem acessar e extrair informações de bancos de dados de arquivos, ou ainda submeter materiais para processamento pela equipe dos bancos, mas não podem adicionar ou alterar as entradas dos bancos diretamente. Esta situação pode mudar. De um ponto de vista prático, a quantidade de dados sendo gerada está aumentando tão rapidamente a ponto de ultrapassar a capacidade que os projetos de armazenamento têm de assimilá-los. Já existe uma tendência de maior envolvimento dos cientistas de bancada na preparação dos dados para submissão aos bancos de armazenamento.

Apesar de haver bons argumentos para o controle exclusivo sobre os arquivos, não existe a necessidade de limitar os meios de acesso a eles – vulgarmente falando, o desenvolvimento de interfaces. Comunidades de usuários especializados podem extrair subconjuntos de dados, ou combinar dados de diferentes fontes e oferecer formas especializadas de acesso. Tais bancos de dados "boutique" dependem dos arquivos primários como fonte de informação, mas re-estruturam a sua própria organização e apresentação. De fato, diferentes bancos de dados secundários podem dividir e manipular a mesma informação de maneiras distintas. Uma extrapolação razoável sugere o conceito de "bancos de dados virtuais" especializados (uma idéia proposta pela primeira vez em 1981), fundamentada nos arquivos, mas fornecendo funções e escopo próprios, direcionados para as necessidades de grupos de pesquisas específicos ou, até mesmo, de cientistas individualmente.

O fluxo da informação na bioinformática

As informações entram no domínio da bioinformática quando um cientista deposita seus resultados experimentais em arquivos de bancos de dados apropriados. A administração do banco de dados organiza e anota os dados, criando uma entrada com conteúdo e formato adequados. A entrada é adicionada ao domínio público do banco de dados. Observe que a divisão do banco de dados em entradas é determinada pela origem dos dados e não pela unidade biológica ou pelo contexto; ou seja, uma entrada corresponde a um conjunto coerente de dados experimentais, geralmente correspondendo, também, a um artigo científico publicado.

Outros projetos de recuperação de informações, em parceria com um banco de dados ou de forma independente, podem integrar uma entrada recém-liberada em seus sistemas individuais. Eles podem selecionar ou reorganizar a estrutura dos dados e fornecer ferramentas para a sua análise.

A reorganização dos dados pode envolver:

- A simples integração da nova entrada a um sistema de pesquisas genéricas ou específicas.
- A extração de subconjuntos dos dados. Exemplos incluem (1) a identificação de genes em uma seqüência de DNA, como em um genoma de bactéria ou um cromossomo eucariótico; e (2) a seleção de um conjunto de seqüências de proteínas não-redundantes, tanto para restringir o espaço de buscas quanto para reduzir a incerteza estatística.
- A derivação de novos tipos de informações a partir dos dados originais. Um exemplo simples: a divulgação de um gene codificador de uma proteína por

um banco de dados contendo seqüências de DNA será seguida pelo surgimento da sua tradução em seqüência de aminoácidos nos bancos de dados de seqüências de proteínas.

- A recombinação dos dados de diferentes formas. Muitos projetos agrupam seqüências ou estruturas de famílias de proteínas homólogas, ou de proteínas que compartilham uma função. Exemplos incluem o banco de dados da protease MEROPS e o Protein Kinase Resource. (Os arquivos em bancos de dados tendem a manter entradas relacionadas separadas para preservar, com transparência, as suas origens.)
- A reanotação dos dados, incluindo o fornecimento de diferentes conjuntos enormes de conexões. A integração pode ser horizontal ou vertical. Isto é, as conexões podem indicar relações com outras entradas do mesmo tipo (por exemplo, correspondências entre genes homólogos em um genoma ou entre genes associados a uma mesma via metabólica). Ou, ainda, essas conexões podem apresentar uma variedade de informações sobre um gene ou proteína (por exemplo, conexões entre um gene e as conseqüências clínicas de suas mutações).

Muitos sítios funcionam como portais entre os arquivos em bancos de dados e as ferramentas computacionais disponíveis para a análise dos dados. A recuperação de informações permite a seleção e a extração de dados a fim de fornecer os componentes de um projeto de pesquisa. Muitos recursos de bioinformática não apenas oferecem a recuperação de informações, mas também facilitam o processamento subseqüente das entradas selecionadas. Um exemplo típico seria recuperar as seqüências de um conjunto de genes homólogos e, após, alinhá-los. O objetivo é fornecer a integração eficiente entre todas as etapas do processamento de dados necessárias para um projeto de pesquisa, por meio de uma conexão robusta entre as ferramentas para armazenamento, recuperação e análise de dados.

Há uma forte tendência para a fusão e a integração das fontes provedoras de dados em bioinformática. Apenas concorrências nacionais ou comerciais parecem estar no caminho de uma extrapolação do que poderá vir a ser, em breve, um único banco mundial de dados. Por causa do risco de este resultado se mostrar muito volumoso e difícil de controlar, a unificação dos bancos de dados deverá ser acompanhada pela fragmentação das vias de acesso.

Organização, anotação e controle de qualidade

As comunidades científica e médica são dependentes da qualidade dos bancos de dados. Índices de qualidade, mesmo que não permitam a correção de erros, podem ajudar a evitar que cheguemos a conclusões erradas.

Entradas de bancos de dados compreendem resultados experimentais brutos e informações suplementares, ou anotações, cada qual com suas próprias margens de erro.

O fator determinante mais importante da qualidade dos dados em si é o estado da arte dos experimentos. Dados antigos têm limitações próprias das técnicas mais antigas; por exemplo, as seqüências de aminoácidos de proteínas eram inicialmente determinadas pelo seqüenciamento de peptídeos, e, agora, são traduzidas a partir de seqüências de DNA (exceto no seqüenciamento parcial por espectrometria de massa; ver Capítulo 6). Uma conseqüência da explosão de dados é que a maior parte dos dados de seqüências é informação nova, gerada pela tecnologia atual, que, na maior parte dos casos, tem um bom desempenho.

As anotações incluem informações sobre a fonte dos dados e a metodologia utilizada para a sua obtenção. Elas identificam os pesquisadores responsáveis e citam as publicações relevantes. Elas fornecem conexões para informações

relacionadas em outros bancos de dados. Em bancos de dados de seqüências, as anotações incluem tabelas de características: listas de segmentos das seqüências que possuem relevância biológica – por exemplo, regiões de uma seqüência de DNA que codificam proteínas. Essas informações aparecem em formatos passíveis de análise por computador, e seus conteúdos podem ser limitados por um vocabulário controlado. Note que a conformidade entre bancos de dados, com relação a um vocabulário controlado e às definições dos termos, é essencial para as operações de recuperação de informação envolvendo interações entre múltiplos bancos de dados, tais como consultas distribuídas.

Antigamente, o registro de uma seqüência de DNA típica era produzido por um único grupo de pesquisa que estudava a relação entre um gene e o seu produto. As anotações se baseavam em dados experimentais gerados localmente e eram escritas por especialistas. Já os projetos de seqüenciamento de genomas completos não oferecem confirmações experimentais da expressão da maior parte dos genes putativos, nem a caracterização de seus produtos. Os administradores de bancos de dados baseiam suas anotações em análises de seqüências por programas computacionais.

A anotação é o ponto mais fraco dos projetos genoma. Sua automação é possível apenas de forma limitada; fazê-la de forma correta ainda exige muito esforço, e os recursos alocados são inadequados. Contudo, a importância de uma anotação precisa não pode ser subestimada. P. Bork frisou que erros na anotação de genes prejudicam a alta qualidade dos próprios dados de seqüências.

O crescimento dos dados genômicos irá permitir um progresso na qualidade da anotação à medida que a precisão dos métodos estatísticos aumenta. Isso permitirá uma *reanotação* mais precisa das entradas. O melhoramento da qualidade da anotação será um fator positivo. Porém, implica uma conseqüência perturbadora; a de que a anotação será instável. Este problema é agravado pela proliferação de sítios na *web* com uma rede de conexões crescente e sobrecarregada. Esses sítios fornecem oportunidades para o desenvolvimento de aplicativos. Mas a *web* é também um vetor para o contágio e a propagação de erros em dados brutos, em dados prematuros que podem ser corrigidos, mas cujas correções não são divulgadas, ou ainda de anotações distintas para uma mesma seqüência.

A única solução possível é um processo distribuído e dinâmico de anotação e correção de erros. Distribuído, pois, em bancos de dados cujos responsáveis não possuem nem tempo nem conhecimento, especialistas terão que atuar como administradores. Dinâmico, pois o progresso na automação da anotação e na identificação e correção de erros permitirá a reanotação dos bancos de dados. Teremos, entretanto, que desistir da idéia cômoda de um banco de dados estável, composto de entradas que estão corretas quando ali depositadas e que permaneçam inalteradas. Os bancos de dados se tornarão uma sopa efervescente de informações, crescendo em tamanho e também – espera-se – em qualidade.

A *World Wide Web*

Todos os leitores já usaram a *World Wide Web* (rede mundial de computadores) para pesquisa de material de referência, para obter notícias, para ter acesso a bancos de dados de biologia molecular, para checar informações particulares de pessoas – amigos, colegas ou celebridades – ou apenas para navegar. A *web* é um meio de contato entre pessoas e entre computadores através de redes. Ela funciona como uma aldeia global completa, contendo o equivalente a bibliotecas, correios, lojas e escolas.

A *web* pode ser imaginada como um gigantesco quadro de avisos multimídia exposto em todo o mundo. Ela contém textos, imagens, filmes e sons. Praticamente, qualquer coisa que possa ser armazenada em um computador pode ser disponibilizada e acessada via *web*. Um exemplo interessante é o sítio sobre a poesia de Walt Whitman (`www.whitmanarchive.org`). A primeira página contém um sumário. O sítio contém os textos impressos de diferentes poemas. Você pode comparar edições diferentes, ter acesso à análise crítica dos poemas e ver alguns deles em versões manuscritas. Existe até uma conexão para um arquivo de áudio, com uma gravação do próprio Whitman lendo parte de um poema.

As conexões contidas em um sítio da *web* podem ser internas ou externas. Conexões internas podem levá-lo a outras partes do texto do documento sendo explorado, ou a imagens, filmes ou sons associados. Conexões externas podem permitir que você se mova verticalmente, *para baixo*, dirigindo-se a documentos mais específicos, ou *para cima*, para documentos mais genéricos (talvez fornecendo desde informações mais básicas até as mais técnicas); ou ainda *horizontalmente*, para documentos similares (outros artigos sobre um mesmo assunto), ou para um *nível mais externo* para diretórios que mostrem que outro material relevante está disponível.

A primeira atitude a tomar, para começar a utilizar a *web* de maneira eficaz, é encontrar os sítios de entrada úteis. Uma vez iniciada uma sessão, as conexões o levarão aonde quer que você deseje ir. Entre os sítios mais importantes, estão as **ferramentas de pesquisa**, como o Google, que indexa toda a *web* e permite a recuperação de dados através de palavras-chave. Você pode entrar com um ou mais termos, tais como "fosforilase", "mudança alostérica", "estrutura cristalina", e o programa que realiza a pesquisa retornará uma lista de conexões para sítios na *web* que contenham esses termos.

Uma vez que você tenha completado uma sessão com sucesso, na próxima vez que você se conectar, as facilidades próprias da memória dos navegadores permitirão que continue, sem problemas, do ponto onde parou. Durante qualquer sessão, quando você se deparar com um documento para o qual gostaria de retornar mais tarde, basta armazenar a conexão em um arquivo de **marcadores de páginas favoritas da *web* (*bookmarks*)** ou **favoritos**. Em uma sessão subseqüente, você pode retornar diretamente para qualquer sítio desta lista, sem necessitar percorrer todo o caminho de conexões que o levaram inicialmente àquele ponto.

Uma *home page* pessoal é um pequeno esboço autobiográfico (com conexões, é claro). Seus colegas poderão ter suas próprias *home pages* que tipicamente contêm nome, afiliação institucional, endereços para correios eletrônico e postal, números de telefone e de fax, uma lista de publicações e de interesses de pesquisas atuais. Não é incomum que *home pages* contenham informações pessoais, tais como passatempos favoritos, fotos da pessoa com o cônjuge e filhos, e até mesmo com seu animal de estimação!

Mas a *web* não é uma via de mão única. Muitos documentos aí contidos incluem formulários, nos quais se pode inserir informações e executar um programa que retorna os resultados dentro da sua sessão. As ferramentas de pesquisa são exemplos comuns. Muitos cálculos em bioinformática são agora realizados por servidores na *web*. Se os cálculos são muito demorados, os resultados podem não ser retornados na mesma sessão, mas enviados por correio eletrônico.

Publicação eletrônica

Estamos em um período notável de transição para a publicação sem papel. Cada vez mais publicações estão surgindo na *web*. Uma revista científica pode publicar na *web* apenas sua lista de conteúdos, ou a lista de conteúdos com os resumos dos artigos, ou mesmo os artigos completos. Muitas publicações institucionais –

boletins e relatórios técnicos – estão disponíveis na *web*. Muitas outras revistas e jornais estão surgindo também. Você pode tentar `http://www.nytimes.com`. Muitas publicações impressas agora contêm referências para conexões na *web* com material suplementar que nunca será publicado em papel. [Enquanto este livro estava sendo encaminhado para publicação, o Google anunciava parcerias com bibliotecas acadêmicas para disponibilizar coleções completas de livros *on-line*. Isso irá criar novos meios de obtenção e transmissão de informações.]

Computadores e a ciência da computação

A bioinformática não seria possível sem os avanços na área de *hardware* e *software* computacionais. Meios de armazenamento rápidos e de alta capacidade são essenciais até para manter os bancos de dados. A obtenção e a análise de informações exigem programas; alguns bastante simples e outros extremamente sofisticados. A distribuição de informações requer as facilidades de redes de computadores e da *World Wide Web*.

A ciência da computação é um campo novo e promissor com o objetivo de fazer o uso mais eficaz do *hardware* da tecnologia da informação. Certas áreas da ciência da computação teórica se relacionam mais diretamente com a bioinformática. Vamos considerá-las em relação a um problema biológico específico: "recuperar de um banco de dados todas as seqüências que são similares a uma seqüência sonda." Uma boa solução para esse problema irá recorrer à ciência da computação para:

- **Análise de algoritmos** *Um algoritmo é uma especificação completa e precisa de um método de resolução de um problema.* Para a recuperação de seqüências similares, precisamos medir a similaridade da seqüência sonda com cada seqüência do banco de dados. É possível fazer muito melhor do que a solução simples de checar cada par de posições em cada justaposição possível, um método que, mesmo sem permitir a inserção de lacunas, exigiria um tempo proporcional ao produto do número de caracteres na seqüência sonda pelo número de caracteres no banco de dados. Uma especialização da ciência da computação, conhecida vulgarmente como *"stringology"*, concentra-se no desenvolvimento de métodos eficientes para este tipo de problema, analisando seus desempenhos efetivos.*
- **Estrutura de dados e recuperação de informação** Como podemos organizar nossos dados para uma resposta eficiente a consultas? Por exemplo, existem meios de indexar ou, de outra maneira, "pré-processar" os dados para tornar mais eficientes nossas pesquisas por similaridade entre seqüências? Como podemos fornecer interfaces que auxiliarão o usuário a conceber e executar essas consultas?
- **Engenharia de *software*** Raramente alguém ainda escreve programas na linguagem natural (de baixo nível) dos computadores. Programadores trabalham com linguagens de alto nível, como C, C++, PERL (Pratical Extraction and Report Language), JAVA ou até mesmo FORTRAN. A escolha da linguagem de programação depende da natureza do algoritmo e da estrutura de dados associada, assim como do uso esperado do programa. Certamente, os *softwares* mais complicados utilizados em bioinformática são escritos por especialistas. O que traz à tona a questão de quanto conhecimento de programação de computadores precisa ter um especialista em bioinformática.

* N. de T. *Stringology* é a ciência que estuda seqüências de caracteres, como a seqüência de uma proteína, DNA ou RNA.

Programação

A programação está para a ciência da computação assim como o assentamento de tijolos está para a arquitetura. Ambos são criativos: um é uma arte; o outro, uma habilidade.

Muitos estudantes de bioinformática perguntam se é imprescindível aprender a escrever programas de computador complicados. Meu conselho (com o qual nem todos da área concordam) é: "Não. A menos que você queira se especializar nisto". Para trabalhar com bioinformática, você deverá se tornar um especialista no uso das ferramentas disponíveis na *web*. Aprender como criar e manter um sítio é essencial. E, é claro, você deverá ter facilidade no uso do sistema operacional do seu computador. Alguma habilidade para escrever *scripts* simples em uma linguagem como PERL fornece uma extensão essencial das funções básicas do sistema operacional.*

Por outro lado, o tamanho dos arquivos de dados e a complexidade crescente das questões que pretendemos responder exigem respeito. Uma programação de alto nível e criativa, nesta área, ficará melhor nas mãos de especialistas bem treinados em ciência da computação. Porém, o uso de programas através de interfaces na *web* muito refinadas, para não dizer vistosas, não fornece qualquer indicação da natureza da atividade envolvida na escrita e na depuração dos programas. Bismarck disse certa vez que "aqueles que adoram lingüiças ou a lei não deveriam ver como ambas são produzidas". Talvez a programação de computadores devesse ser incluída nesta lista.

Eu recomendo o aprendizado de algumas habilidades básicas em PERL, ou em uma das linguagens relacionadas, Python ou Ruby. Essas linguagens permitem o desenvolvimento de ferramentas poderosas. Elas facilitam bastante a execução de muitas tarefas simples e úteis, e estão disponíveis na maioria dos sistemas de computadores.

O quanto se deve aprender de PERL para que seja útil em bioinformática? Muitas instituições ministram cursos. Aprender com colegas é válido, dependendo da sua aptidão e da paciência de seus colegas. Livros também estão disponíveis. Um meio bastante útil é encontrar tutoriais na *web* – procure em um sítio de pesquisas por "tutorial PERL" e você encontrará muitos sítios úteis que lhe ensinarão o básico.

E, é claro, utilize-o o máximo que você puder. Este livro não vai lhe ensinar PERL, mas lhe dará oportunidades de praticar o que você aprendeu em outros lugares. Se suas ambições quanto à programação forem além das tarefas simples, visite o projeto Bioperl, uma fonte de programas e bibliotecas em PERL aplicada à bioinformática (ver `http://bio.perl.org/`) e disponível gratuitamente.

Exemplos de programas *simples* em PERL são descritos neste livro. O poder da linguagem PERL em lidar com caracteres e/ou seqüência de caracteres a torna conveniente na resolução de problemas que envolvem seqüências biológicas. Aqui está um programa PERL bastante simples para a tradução de seqüências de nucleotídeos em seqüência de aminoácidos de acordo com o código genético padrão. A primeira linha, `#!/usr/bin/perl`, é um sinal para o sistema operacional UNIX (ou LINUX) de que o que segue é um programa PERL. Dentro do programa, todo texto iniciado com #, até o final da linha em que aparece, é

* N. de T. *Scripts* são um conjunto de comandos, que podem ser escritos em diversas linguagens como PERL, Python, etc., armazenados em um arquivo-texto, que são executados seqüencialmente.

Exemplo de PERL 1.1 Tradução de uma seqüência de DNA em uma seqüência de aminoácidos utilizando o código genético padrão

```
#!/usr/bin/perl
#translate.pl -- translate nucleic acid sequence to protein sequence
#               according to standard genetic code

#   set up table of standard genetic code

%standardgeneticcode = (
 "ttt"=> "Phe",   "tct"=> "Ser", "tat"=> "Tir",   "tgt"=> "Cis",
 "ttc"=> "Phe",   "tcc"=> "Ser", "tac"=> "Tir",   "tgc"=> "Cis",
 "tta"=> "Leu",   "tca"=> "Ser", "taa"=> "TER",   "tga"=> "TER",
 "ttg"=> "Leu",   "tcg"=> "Ser", "tag"=> "TER",   "tgg"=> "Trp",
 "ctt"=> "Leu",   "cct"=> "Pro", "cat"=> "His",   "cgt"=> "Arg",
 "ctc"=> "Leu",   "ccc"=> "Pro", "cac"=> "His",   "cgc"=> "Arg",
 "cta"=> "Leu",   "cca"=> "Pro", "caa"=> "Gln",   "cga"=> "Arg",
 "ctg"=> "Leu",   "ccg"=> "Pro", "cag"=> "Gln",   "cgg"=> "Arg",
 "att"=> "Ile",   "act"=> "Tre", "aat"=> "Asn",   "agt"=> "Ser",
 "atc"=> "Ile",   "acc"=> "Tre", "aac"=> "Asn",   "agc"=> "Ser",
 "ata"=> "Ile",   "aca"=> "Tre", "aaa"=> "Lis",   "aga"=> "Arg",
 "atg"=> "Met",   "acg"=> "Tre", "aag"=> "Lis",   "agg"=> "Arg",
 "gtt"=> "Val",   "gct"=> "Ala", "gat"=> "Asp",   "ggt"=> "Gli",
 "gtc"=> "Val",   "gcc"=> "Ala", "gac"=> "Asp",   "ggc"=> "Gli",
 "gta"=> "Val",   "gca"=> "Ala", "gaa"=> "Glu",   "gga"=> "Gli",
 "gtg"=> "Val",   "gcg"=> "Ala", "gag"=> "Glu",   "ggg"=> "Gli"
);


#   process input data

while ($line = <DATA>) {                                  # read in line of input
    print "$line";                                        # transcribe to output
    chop();                                               # remove end-of-line character
    @triplets = unpack("a3" x (length($line)/3),$line);# pull out successive triplets
    foreach $codon (@triplets) {                          # loop over triplets
        print "$standardgeneticcode{$codon}";             # print out translation of each
    }                                                     # end loop on triplets
    print "\n\n";                                         # skip line on output
}                                                         # end loop on input lines

#   what follows is input data

__END__
atgcatccctttaat
tctgtctga
```

Executando este programa com os dados de entrada fornecidos, obtém-se como resultado:

```
atgcatccctttaat
MetHisProPheAsn

tctgtctga
SerValTER
```

apenas um comentário. A linha `__END__` indica o término do programa e que a informação seguinte são os dados de entrada.

Mesmo esse programa simples mostra muitas características da linguagem PERL. O arquivo contém as informações básicas (a tabela de tradução do código genético), instruções que dizem ao computador o que fazer e os dados de entrada (que aparecem após a linha `__END__`). Os comentários explicam brevemente as seções do programa e descrevem o efeito de cada instrução.

O programa é estruturado em blocos mantidos entre chaves: { . . . }, as quais são úteis no controle do fluxo de execução. Dentro dos blocos, instruções individuais (cada uma terminando com um ;) são executadas na ordem em que aparecem. O bloco externo, que não está entre chaves, representa um laço:

```
while ($line = <DATA>) {
    ...
}
```

Aqui, <DATA> se refere às linhas dos dados de entrada (que aparecem após __END__). O bloco é executado uma vez para cada linha dos dados de entrada; ou seja, `while` indica enquanto houver linhas nos dados de entrada que ainda não foram lidas.

Três tipos de estruturas de dados aparecem no programa. A linha dos dados de entrada, referida como `$line`, é uma simples **seqüência de caracteres**. Ela é dividida em uma **matriz** ou vetor de tripletos de nucleotídeos. Uma matriz armazena diversos itens em uma ordem linear, e itens individuais dos dados podem ser recuperados a partir de suas posições na matriz. Para facilitar a consulta a um aminoácido codificado por qualquer tripleto, o código genético é armazenado como uma **matriz associativa**. Uma matriz associativa, ou uma tabela de busca, é uma generalização de uma matriz simples ou seqüencial. Enquanto os elementos de uma matriz simples são indexados por inteiros consecutivos, os elementos de uma matriz associativa são indexados por qualquer seqüência de caracteres, neste caso, os 64 tripletos. Utilizamos os tripletos de entrada *na ordem em que aparecem* na seqüência nucleotídica, mas precisamos acessar os elementos da tabela do código genético *em uma ordem arbitrária*, determinada pela sucessão de tripletos. Uma matriz simples ou um vetor de seqüência de caracteres são apropriados para o processamento de tripletos sucessivos, e a matriz associativa é apropriada para consulta aos aminoácidos correspondentes.

Exemplo de PERL 1.2 Montagem de fragmentos com sobreposições

Este é outro programa PERL que ilustra aspectos adicionais da linguagem.* Esse programa reagrupa a frase:

> All the world's a stage,
> And all the men and women merely players;
> They have their exits and their entrances,
> And one man in his time plays many parts.

após ela ter sido dividida em fragmentos aleatórios com sobreposições (\n nos fragmentos representa o fim da linha na frase original):

> the men and women merely players;\n
> one man in his time
> All the world's
> their entrances,\nand one man
> stage,\nAnd all the men and women
> They have their exits and their entrances,\n
> world's a stage,\nAnd all
> their entrances,\nand one man
> in his time plays many parts.
> merely players;\nThey have

Este tipo de cálculo é importante na montagem de seqüências de DNA a partir de fragmentos com sobreposições (ver Problemas 1.5 e 1.6).

*Esta seção pode ser pulada em uma primeira leitura.

→

Exemplo de PERL 1.2 (*continuação*)

```
#!/usr/bin/perl
#assemble.pl -- assemble overlapping fragments of strings
#  input of fragments
while ($line = <DATA>) {                    #   read in fragments, 1 per line
   chop($line);                             #   remove trailing carriage return
   push(@fragments,$line);                  #   copy each fragment into array
}
#  now array  @fragments  contains fragments
#  we need two relationships between fragments:
#  (1) which fragment shares no prefix with suffix of another fragment
#      * This tells us which fragment comes first
#  (2) which fragment shares longest suffix with a prefix of another
#      * This tells us which fragment  follows  any fragment
#  First set array of prefixes to the default value   "noprefixfound".
#      Later, change this default value when a prefix is found.
#      The one fragment that retains the default value must be come first.
#  Then loop over pairs of fragments to determine maximal overlap.
#      This determines successor of each fragment
#      Note in passing that if a fragment has a successor then the
#          successor must have a prefix
foreach $i (@fragments) {                   #   initially set  prefix of each fragment
    $prefix{$i} = "noprefixfound";          #       to  "noprefixfound"
    }                                       #   this will be overwritten when a prefix is found
#  for each pair, find longest overlap of suffix of one with prefix of the other
#       This tells us which fragment FOLLOWS any fragment
foreach $i (@fragments) {                   #   loop over fragments
    $longestsuffix = "";                    #   initialize longest suffix to null
    foreach $j (@fragments) {               #   loop over fragment pairs
        unless ($i eq $j) {                 #   don't check fragment against itself
            $combine = $i . "XXX" . $j; #   concatenate fragments, with fence XXX
            $combine =~ /([\S ]{2,})XXX\1/;   #   check for repeated sequence
            if (length($1) > length($longestsuffix)) {    # keep longest overlap
                $longestsuffix = $1;    #   retain longest suffix
                $successor{$i} = $j;    #   record that $j follows $i
            }
        }
    }
    $prefix{$successor{$i}} = "found";  #  if $j follows $i then $j must have a prefix
}

foreach (@fragments) {                      #  find fragment that has no prefix; that's the start
    if ($prefix{$_} eq "noprefixfound") {$outstring = $_;}
}
$test = $outstring;                         #   start with fragment without prefix
while ($successor{$test}) {                 #   append fragments in order
     $test = $successor{$test};             #   choose next fragment
     $outstring = $outstring . "XXX" . $test;  # append to string
     $outstring =~ s/([\S ]+)XXX\1/\1/; #   remove overlapping segment
}
$outstring =~ s/\\n/\n/g;                   #   change signal \n to real carriage return
print "$outstring\n";                       #   print final result
__END__
the men and women merely players;\n
one man in his time
All the world's
their entrances,\nand one man
stage,\nAnd all the men and women
They have their exits and their entrances,\n
world's a stage,\nAnd all
their entrances,\nand one man
in his time plays many parts.
merely players;\nThey have
```

Classificação e nomenclatura biológica

Vamos retornar ao século XVIII, quando a vida acadêmica era mais simples, pelo menos em alguns aspectos.

A nomenclatura biológica se baseia na idéia de que os organismos vivos são divididos em unidades denominadas espécies – grupos de organismos similares com um reservatório genético comum. (Por que organismos vivos deveriam ser "quantizados" em espécies *discretas* é uma questão bastante complicada.) Linnaeus, um naturalista sueco, classificou os organismos vivos de acordo com uma hierarquia: Reino, Filo, Classe, Ordem, Família, Gênero e Espécie (ver Quadro). Taxonomistas modernos adicionaram mais níveis a esta classificação. Para a identificação, geralmente é suficiente especificar o binômio Gênero e Espécie; por exemplo, *Homo sapiens* para humanos e *Drosophila melanogaster* para a mosca-das-frutas. Cada binômio descreve uma única espécie, que também pode ser conhecida por um ou mais nomes comuns; por exemplo, *Bos taurus* = vaca. Certamente, a maioria das espécies não possui nomes comuns.

Classificações do ser humano e da mosca-das-frutas

	Ser humano	Mosca-das-frutas
Reino	Animalia	Animalia
Filo	Chordata	Arthropoda
Classe	Mammalia	Insecta
Ordem	Primata	Diptera
Família	Hominidae	Drosophilidae
Gênero	*Homo*	*Drosophila*
Espécie	*sapiens*	*melanogaster*

Originalmente, o sistema de Linnaeus era apenas uma classificação baseada nas similaridades observadas. Com a descoberta da evolução, percebeu-se que esse sistema refletia basicamente a ancestralidade biológica. A questão sobre quais semelhanças realmente refletem uma ancestralidade comum deve agora ser encarada. Características derivadas de um ancestral comum são chamadas de homólogas; por exemplo, as asas de uma águia e os braços de um ser humano. Outras características aparentemente similares podem ter surgido de forma independente por evolução convergente; por exemplo, as asas de uma águia e as asas de uma abelha. De modo oposto, características realmente homólogas podem ter divergido para se tornarem muito diferentes em estrutura e função. Os ossos do ouvido médio dos humanos são homólogos aos ossos das mandíbulas dos peixes primitivos; nossas trompas de Eustáquio são homólogas às guelras dos peixes. Na maioria dos casos, os especialistas conseguem distinguir as homologias genuínas das similaridades resultantes de evolução convergente.

A análise de seqüências fornece a evidência mais clara das relações entre as espécies. O sistema funciona bem para os organismos superiores, para os quais

a análise de seqüências e as ferramentas clássicas da anatomia comparativa, paleontologia e embriologia fornecem, normalmente, dados consistentes. A classificação de microrganismos é mais difícil, em parte porque a seleção de características para fundamentar as suas classificações é menos óbvia, e em parte pela grande quantidade de transferência gênica lateral que ameaça mudar completamente o cenário.

Os RNAs ribossomais têm a característica essencial de estarem presentes em todos os organismos, com o grau ideal de divergência. (Graus de divergência e de parentesco muito elevados ou muito pequenos tornam-se invisíveis, ou seja, são difíceis de serem estimados.)

Com base nos RNAs ribossomais 16S, C. Woese dividiu os organismos vivos em três domínios fundamentais (um nível *acima* de Reino na hierarquia): Bacteria, Archaea e Eukarya (ver Figura 1.2). Os domínios Bacteria e Archaea são constituídos de procariotos; suas células não contêm núcleo. O domínio Bacteria inclui os microrganismos tipicamente responsáveis por muitas doenças infecciosas e, é claro, *Escherichia coli*, o modelo principal da biologia molecular. O domínio Archaea compreende os termófilos e halófilos extremos, os redutores de sulfato e os metanogênicos. Nós pertencemos ao domínio Eukarya – organismos cujas células contêm núcleo, incluindo a levedura e todos os organismos multicelulares.

Um levantamento das espécies com genomas seqüenciados destaca as bactérias, por causa de sua importância clínica e pela facilidade relativa do seqüenciamento dos genomas de procariotos. Contudo, fundamentalmente, temos mais a aprender sobre nós mesmos a partir de estudos com archaeas do que com bactérias. Pois, sem considerar as diferenças óbvias no estilo de vida, e a ausência de um núcleo, as archaeas são, em alguns pontos, mais próximas dos eucariotos do que das bactérias em nível molecular. É provável também que as archaeas sejam os organismos vivos mais próximos da raiz na árvore da vida.

A Figura 1.2 mostra os níveis mais básicos da árvore da vida. O ramo do domínio Eukarya inclui animais, plantas e fungos. No final desse ramo estão os metazoários (organismos multicelulares – Figura 1.3). Nós e nossos parentes mais próximos somos deuterostômios (Figura 1.4).

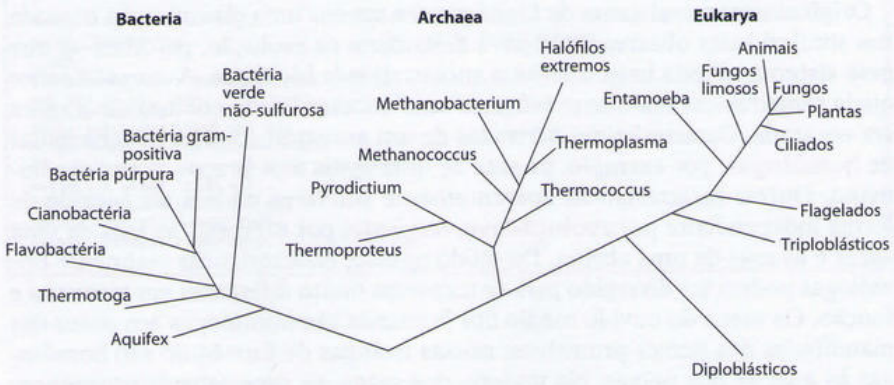


Figura 1.2 Divisões principais dos organismos vivos, derivadas por C. Woese com base nas seqüências de RNAs ribossomais 16S.

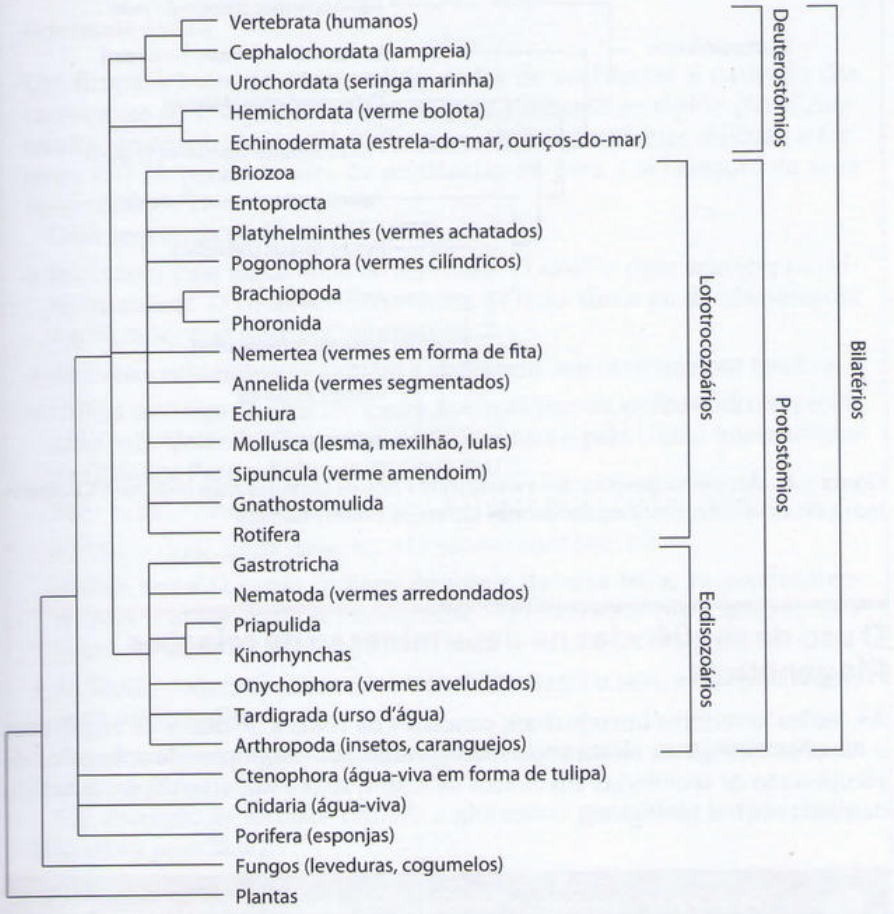


Figura 1.3 Árvore filogenética dos metazoários (animais multicelulares). Os bilatérios incluem todos os animais que compartilham simetria lateral (direita/esquerda) no plano corporal. Protostômios e deuterostômios são duas linhagens principais separadas nos estágios iniciais da evolução há aproximadamente 670 milhões de anos. Ambos mostram padrões bastante diferentes de desenvolvimento embrionário, incluindo diferentes padrões iniciais de divisão celular denominada clivagem, orientações opostas do intestino completo em relação à invaginação inicial da blástula, e a origem do esqueleto a partir da mesoderme (deuterostômios) ou ectoderme (protostômios). Os protostômios compreendem dois subgrupos diferenciados pelas seqüências do RNA 18S (da subunidade ribossomal menor) e do gene HOX. Morfologicamente, os ecdisozoários possuem uma cutícula protetora – uma camada externa rígida composta de material orgânico. Os lofotrocozoários têm corpos moles. (Baseado em Adouette, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B. & de Rosa, R. (2000), The new animal phylogeny: Reliability and implications, *Proceedings of National Academy of Sciences USA*, **97**:4453-4456.)

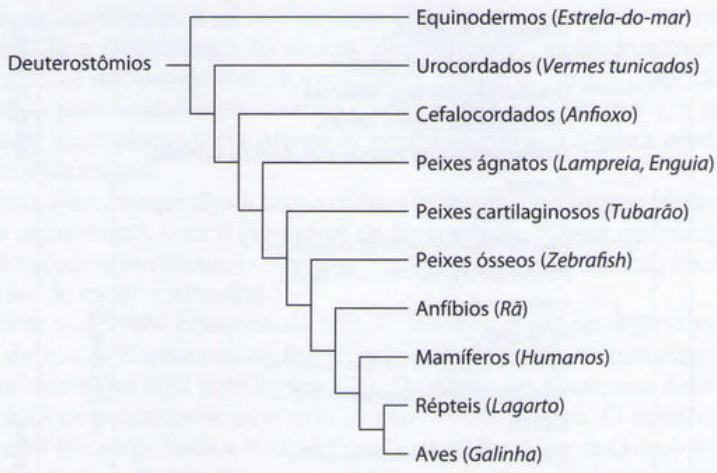


Figura 1.4 Árvore filogenética dos vertebrados e nossos parentes mais próximos. Cordados, incluindo os vertebrados, e equinodermos são todos deuterostômios.

O uso de seqüências na determinação de relações filogenéticas

As seções anteriores introduziram conceitos de bancos de dados de seqüências e relações biológicas. Nesta seção, são apresentados exemplos de aplicações de recuperação de seqüências em bancos de dados, suas comparações, até a análise de suas relações biológicas.

Estudo de Caso 1.1: Obtenha a seqüência de aminoácidos da ribonuclease pancreática de cavalo (*horse pancreatic ribonuclease*)

Use o servidor ExPASy do Swiss Institute for Bioinformatics. A sua URL é `http://expasy.org/cgi-bin/sprot-search-ful`. Digite as palavras-chave `horse pancreatic ribonuclease` e então pressione a tecla `ENTER`. Selecione `RNP_HORSE` e então o `formato FASTA` (ver Quadro O formato FASTA). O resultado será o seguinte (após truncar a primeira linha):

```
>sp|P00674|RNP_HORSE RIBONUCLEASE PANCREATIC (EC 3.1.27.5) (RNASE 1)...
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIVACEGNPYVPVHFDASVEVST
```

o qual pode ser selecionado e utilizado em outros programas.*

Por exemplo, poderíamos selecionar diversas seqüências e alinhá-las (ver Quadro Alinhamento de seqüências). As análises de padrões de similaridade de seqüências alinhadas são bastante úteis na avaliação de suas relações de parentesco.

* N. de T. Como a bioinformática é uma disciplina bastante dinâmica e os bancos de dados e as interfaces estão sendo constantemente melhorados e atualizados, pode ocorrer de o resultado apresentado neste livro não ser exatamente o que se obtém na data atual da sua pesquisa. As diferenças devem estar mais relacionadas ao formato do que ao conteúdo.

O formato FASTA

Um formato bastante comum para dados de seqüências é derivado das convenções do FASTA, um programa para alinhamento rápido (*FAST Alignment*), desenvolvido por W. R. Pearson. Muitos programas utilizam o formato FASTA para a leitura de seqüências ou para a informação de seus resultados.

Uma seqüência no formato FASTA:

- Inicia com uma única linha de descrição. O sinal > deve aparecer na primeira coluna. O conteúdo do restante da linha título ou de identificação é arbitrário, mas deve ser informativo.
- As linhas subseqüentes contêm a seqüência, um caractere por resíduo.
- Utiliza o código de uma letra para nucleotídeos ou aminoácidos especificado pela União Internacional de Bioquímica e pela União Internacional de Química Pura e Aplicada (IUB/IUPAC).

  Ver: `http://www.chem.qmw.ac.uk/iupac/misc/naabb.html`

  e `http://www.chem.qmw.ac.uk/iupac/AminoAcid/`

  Utilize Sec e U como códigos de três e de uma letra, respectivamente, para o aminoácido selenocisteína: `http://www.chem.qmw.ac.uk/iubmb/newsletter/1999/item3.html`
- As linhas podem ter comprimentos diferentes; ou seja, a margem direita pode ser irregular.
- A maioria dos programas aceitará letras minúsculas para seqüências de aminoácidos.

Um exemplo de formato FASTA: a glutationa peroxidase bovina (*bovine glutathione peroxidase*).

```
>gi|121664|sp|P00435|GSHC_BOVIN GLUTATHIONE PEROXIDASE
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEPFNLSSLRGKVLLIENVASLUGTTVRDYTQMNDLQRRLG
PRGLVVLGFPCNQFGHQENAKNEEILNCLKYVRPGGGFEPNFMLFEKCEVNGEKAHPLFAFLREVLPTPS
DDATALMTDPKFITWSPVCRNDVSWNFEKFLVGPDGVPVRRYSRRFLTIDIEPDIETLLSQGASA
```

A linha título ou de identificação contém as seguintes informações:

O sinal > obrigatório na coluna 1.

`gi|121664` é o seu *número geninfo*, um identificador atribuído pelo US National Center for Biotechnology Information (NCBI) para cada seqüência no seu banco de dados ENTREZ. O NCBI coleta seqüências a partir de várias fontes, incluindo coleções de arquivos de dados primários e pedidos de patentes. Seus números "gi" fornecem um identificador, do tipo "guarda-chuva", comum e consistente, para as seqüências, sobrepondo as diferentes convenções dos bancos de dados fonte. Quando um banco de dados fonte atualiza uma entrada, o NCBI cria uma nova entrada com um novo número "gi" se a atualização alterar a seqüência, mas apenas atualiza e mantém sua entrada quando as alterações afetarem apenas as informações que não concernem à seqüência, tal como referências na literatura.

`sp|P00435` indica que o banco de dados fonte é o SWISS-PROT, e que o número de acesso da entrada no SWISS-PROT é P00435.

`GSHC_BOVIN GLUTATHIONE PEROXIDASE` é o identificador da seqüência e da espécie (GSHC_BOVIN), no SWISS-PROT, seguido pelo nome da molécula.

Alinhamento de seqüências

O alinhamento de seqüências é a atribuição de correspondências entre pares de resíduos. Nós queremos encontrar:

- *Um alinhamento Global*: alinhar todos os caracteres de uma seqüência com todos os caracteres da outra seqüência.

```
And.--so,.from.hour.to.hour,.we.ripe.and.ripe
||||    ||||||||||||||||||||||||    |||||||
And.then,.from.hour.to.hour,.we.rot-.and.rot-
```

  Este exemplo ilustra malpareamentos, inserções e deleções.

- *Um alinhamento Local*: encontrar uma região em uma seqüência que se alinha a uma região de outra seqüência.

```
   My.care.is.loss.of.care,.by.old.care.done,
     |||||||||     |||||||||||||    ||||||  ||
Your.care.is.gain.of.care,.by.new.care.won
```

  Para o alinhamento local, os caracteres que não pareiam nas extremidades não são tratados como lacunas. Além de malpareamentos, vistos neste exemplo, inserções e deleções também são possíveis na região alinhada.

- *Um alinhamento de Motivos*: encontrar alinhamentos de uma seqüência pequena com uma ou mais regiões internas de uma seqüência longa.

```
    match
    |||||
The match is made; she seals it with a curtsy.
```

  Pode-se permitir o malpareamento de alguns caracteres:

```
          match
           ||||
  for the watch to babble and to talk is most tolerable
or:   match                                 match
        |||                                 ||  |
  And witch the world with noble horsemanship.
```

ou inserções e/ou deleções:

```
          mat--ch     mat-ch
          ||    |     ||  |
Fear not, Macbeth; no man that's born of woman
Shall e'er have power upon thee.
```

- *Um alinhamento Múltiplo*: um alinhamento simultâneo de muitas seqüências.

```
no.sooner.---met.---------but.they.-look'd
no.sooner.look'd.---------but.they.-lo-v'd
no.sooner.lo-v'd.---------but.they.-sigh'd
no.sooner.sigh'd.---------but.they.--asked.one.another.the.reason
no.sooner.knew.the.reason.but.they.------------sought.the.remedy
no.sooner.                .but.they.
```

A última linha mostra os caracteres conservados em todas as seqüências do alinhamento múltiplo. Ver Capítulo 4 para uma discussão mais detalhada sobre alinhamentos.

Estudo de Caso 1.2

Determine, a partir das seqüências da ribonuclease pancreática do cavalo (*Equus caballus*), da baleia-branca (*Balaenoptera acutorostrata*) e do canguru-vermelho (*Macropus rufus*), quais destas espécies são mais proximamente relacionadas.

Sabendo-se que o cavalo e a baleia são animais placentários e que o canguru é um marsupial, espera-se que o cavalo e a baleia sejam as espécies mais relacionadas. Obtendo as três seqüências como no exemplo anterior e as inserindo como no formato abaixo:

```
>RNP_HORSE
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIVACEGNPYVPVHFDASVEVST
>RNP_BALAC
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHES
LEDVKAVCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTS
QKEKHIIVACEGNPYVPVHFDNSV
>RNP_MACRU
ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPK
SVVDAVCHQENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSN
LNKQIIVACEGQYVPVHFDAYV
```

no programa CLUSTAL-W de alinhamento múltiplo de seqüências
http://www.ebi.ac.uk/clustalw/
(ou, alternativamente, T-coffee:
http://www.ch.embnet.org/software/TCoffee.html)
obtém-se o seguinte resultado:

```
CLUSTAL W (1.8) mutiple sequence alignment

RNP_HORSE    KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAICLQ 60
RNP_BALAC    RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKAVCSQ 60
RNP_MACRU    -ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPKSVVDAVCHQ 59
              *:** **:*****:  :......*** **  *.**.* ***:***:**.   *.*:* *

RNP_HORSE    KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF 120
RNP_BALAC    KNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPVHF 120
RNP_MACRU    ENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSNLNKQIIVACEG-QYVPVHF 118
             :*: ****::***:*.* : **:** *..****** *:**: :::******** ******

RNP_HORSE    DASVEVST 128
RNP_BALAC    DNSV---- 124
RNP_MACRU    DAYV---- 122
             *  *
```

Nesta tabela, um * sob as seqüências indica uma posição conservada (o resíduo é idêntico em todas as seqüências), e : ou . indicam posições em que todas as seqüências contêm resíduos de características físico-químicas muito similares (:), ou com pouca similaridade (.).

Segmentos consideráveis das seqüências são idênticos. Há várias substituições, mas apenas uma deleção interna. Comparando as seqüências *par a par*, o número de resíduos idênticos (não é o mesmo que contar *s no alinhamento múltiplo acima) entre os pares desse alinhamento é:

Número de resíduos idênticos no alinhamento das seqüências de ribonuclease A (de um total de 122 a 128 resíduos)

Cavalo	e	Baleia-branca	95
Baleia-branca	e	Canguru-vermelho	82
Cavalo	e	Canguru-vermelho	75

→

O cavalo e a baleia compartilham o maior número de resíduos idênticos. Este resultado parece expressivo e, portanto, confirma nossas expectativas. *Atenção: Ou seria a lógica exatamente o contrário?*

Estudo de Caso 1.3

Vamos tentar um exemplo mais complicado:

Os dois gêneros vivos de elefantes são representados pelo elefante africano (*Loxodonta africana*) e indiano (*Elephas maximus*). Foi possível seqüenciar o citocromo *b* de mitocôndrias de uma espécime de mamute lanoso siberiano (*Mammuthus primigenius*), conservado no subsolo permanentemente congelado do Ártico. Com qual elefante moderno o mamute está mais relacionado?

Obtendo as seqüências e executando o CLUSTAL-W, chega-se ao seguinte resultado:

```
Indian Elephant    MTHTRKFHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
Siberian Mammoth   MTHIRKSHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
African Elephant   MTHIRKSHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60

                   *** ** ***:**:**************************************************

Indian Elephant    TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
Siberian Mammoth   TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
African Elephant   TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
                   ************************************************************

Indian Elephant    LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA 180
Siberian Mammoth   LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTDLVEWIWGGFSVDKATLNRFFA 180
African Elephant   LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA 180
                   ************************************:***********************

Indian Elephant    FHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL 240
Siberian Mammoth   LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLL 240
African Elephant   LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL 240
                   :********:************************************************:**

Indian Elephant    LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSILI 300
Siberian Mammoth   LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI 300
African Elephant   LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI 300
                   ************************************************:*****

Indian Elephant    LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIIIGQMASILYFS 360
Siberian Mammoth   LGIMPLLHTSKHRSMMLRPLSQVLFWTLATDLLMLTWIGSQPVEYPYIIIGQMASILYFS 360
African Elephant   LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIGQMASILYFS 360
                   **:************************: *** **********:***************

Indian Elephant    IILAFLPIAGMIENYLIK 378
Siberian Mammoth   IILAFLPIAGMIENYLIK 378
African Elephant   IILAFLPIAGVIENYLIK 378
                   **********:*******
```

Enquanto as seqüências do mamute e do elefante africano apresentam 8 malpareamentos, as seqüências do mamute e do elefante indiano apresentam 14. Parece que o mamute está mais proximamente relacionado aos elefantes africanos. Porém, este resultado é menos satisfatório do que o anterior. Há menos diferenças entre as seqüências. Elas são significativas? (Neste caso, é mais difícil decidir se as diferenças são significativas porque não temos uma idéia preconcebida sobre qual deveria ser a resposta.)

Este exemplo levanta algumas questões:

(1) Nós "sabemos" que os elefantes africano e indiano e o mamute devem ser parentes próximos; basta olhar para eles. Mas podemos afirmar, *a partir apenas destas seqüências*, que elas pertencem a espécies proximamente relacionadas?

(2) Dado que as diferenças são poucas, elas representam de fato uma seleção natural ou apenas erro ou flutuação aleatórios? Precisamos de critérios estatísticos sensíveis para julgar a significância dessas similaridades e diferenças. Seria útil verificar os próprios genes e checar a razão entre o número de substituições não-sinônimas e sinônimas. (Uma substituição sinônima é uma alteração na seqüência de ácidos nucléicos que não altera a seqüência da proteína que ela codifica. Ver páginas 27-28.) Uma razão alta entre substituições não-sinônimas e sinônimas sugere divergência sob pressão seletiva.

Como embasamento para tais questões, vamos destacar a diferença principal entre **similaridade** e **homologia**. *Similaridade* é a observação ou mensuração de semelhança e diferença, independentemente da origem da semelhança. *Homologia* significa, especificamente, que as seqüências e os organismos nos quais ocorrem descendem de um ancestral comum, com a implicação de que as similaridades são características ancestrais compartilhadas. A similaridade de seqüências (ou de características biológicas macroscópicas) é visível nos dados coletáveis *atualmente*, e não envolve hipóteses históricas. Ao contrário, afirmações sobre homologia envolve eventos históricos, os quais, na maioria das vezes, não são observáveis. A homologia tem de ser uma *inferência* a partir da observação de similaridades. Apenas em alguns casos especiais a homologia é observável diretamente; por exemplo, em árvores genealógicas de famílias que apresentam fenótipos incomuns, tal como o lábio de Hapsburg, ou em populações de laboratório, ou em estudos clínicos que monitoram, em nível de seqüências, o curso de infecções virais em pacientes individuais.

A afirmação de que os citocromos *b* dos elefantes africano e indiano e de mamute são homólogos *significa* que existiu um ancestral comum, provavelmente contendo um único tipo de citocromo *b* que, por meio de mutações alternativas, originou as proteínas de mamute e dos elefantes modernos. O alto grau de similaridade entre estas seqüências justifica a conclusão de que elas são homólogas, ou pode haver outras explicações?

- Pode ser que um citocromo *b* funcional *exija* tantos aminoácidos conservados que os citocromos *b* de todos os animais são tão similares uns aos outros quanto são as proteínas de mamute e dos elefantes. Podemos testar esta hipótese verificando a seqüência do citocromo *b* de outras espécies. O resultado é que os citocromos *b* das outras espécies diferem substancialmente dos de elefante e de mamute.
- Uma segunda possibilidade é que existam necessidades fisiológicas especiais para o funcionamento adequado do citocromo *b* em animais com a forma e o tamanho de um elefante, de maneira que as três seqüências de citocromos *b* analisadas surgiram de ancestrais distintos e, submetidas à mesma pressão seletiva, tenham se tornado similares. (Lembre-se de que estamos perguntando sobre o que pode ser deduzido a partir apenas das seqüências do citocromo *b*.)
- O mamute pode estar mais relacionado com o elefante indiano e, desde a época do último ancestral comum, a seqüência do citocromo *b* do elefante indiano evoluiu mais rápido do que a do elefante africano ou do mamute, acumulando mais mutações.

- Existe ainda a possibilidade de uma quarta hipótese: a de que todos os ancestrais comuns de elefantes e mamutes tinham citocromos *b* muito diferentes, mas que os elefantes e os mamutes vivos adquiriram um gene comum, de um organismo não-relacionado, transferido por um vírus.

Suponha, contudo, que concluamos que a similaridade entre as seqüências de elefante e mamute é alta o suficiente para implicar homologia; então, o que dizer das seqüências de ribonuclease do exemplo anterior? A diferença *maior* entre as ribonucleases pancreáticas de cavalo, baleia e canguru é uma evidência de que elas *não* sejam homólogas?

Como podemos responder a estas questões? Especialistas tomaram cuidados extras na calibração das similaridades e divergências entre muitas proteínas, de muitas espécies, cujas relações taxonômicas foram estudadas por métodos clássicos. No exemplo das ribonucleases pancreáticas, a conclusão de que a similaridade implica homologia é justificada. A questão de qual espécie é mais proximamente relacionada ao mamute, se o elefante africano ou indiano, só foi decidida recentemente em favor dos elefantes africanos. A análise de similaridades entre seqüências em genomas e proteínas está tão bem estabelecida que pode ser considerada o método mais eficaz na determinação de relações filogenéticas, mesmo que em alguns casos os resultados não sejam significativos ou, em outros, nem mesmo levem à resposta correta. Existem muitos dados disponíveis, assim como ferramentas eficientes para se obter a informação necessária para o tratamento de questões específicas, e ferramentas de análise bastante robustas. Nenhum desses benefícios, porém, substitui a necessidade de uma avaliação científica criteriosa.

O uso de SINES e LINES na derivação de relações filogenéticas

Os principais problemas na inferência de filogenias por meio da comparação de seqüências de genes e de proteínas são (1) a grande variação da similaridade, que pode ficar bem abaixo da significância estatística, e (2) os efeitos de diferentes taxas de evolução em ramos distintos da árvore evolucionária. Em muitos casos, mesmo que a similaridade entre seqüências estabeleça relações confiáveis, pode ser impossível determinar a *ordem* em que os grupos de táxons (ou *taxa*, em latim) se separaram. O sonho dos especialistas em filogenética – características do tipo "tudo ou nada", cuja manifestação é irreversível de forma que a ordem de ramificação dos eventos pode ser decidida – é, em alguns casos, proporcionado por seqüências não-codificadoras nos genomas.

SINES e LINES (para Short Interspersed Nuclear ElementS e Long Interspersed Nuclear ElementS) são seqüências não-codificadoras, repetitivas, que representam grandes frações do genoma de eucariotos – pelo menos 30% do DNA cromossômico humano e mais de 50% em alguns genomas de plantas superiores. Tipicamente, os SINES possuem entre 70 e 500 pares de base, e estima-se que até um milhão dessas seqüências estejam espalhadas no genoma humano. Os LINES podem ter até 7.000 pares de base, e o seu número de cópias no genoma humano pode chegar a centenas de milhares de cópias. Os SINES são incorporados ao genoma pela transcrição reversa de RNA. A maioria dos SINES contém uma porção 5' homóloga ao tRNA, uma região central não relacionada ao tRNA, e uma porção 3' rica em AT.

Características dos SINES que os tornam úteis no estudo de filogenias incluem:

- Um SINE está presente ou ausente. A presença de um SINE em uma posição específica é uma propriedade que não acarreta nenhuma mensuração variável e complicada de similaridade.

- Os SINES são inseridos aleatoriamente na porção não-codificadora de um genoma. Portanto, a presença de SINES similares no mesmo *locus*, em duas espécies, implica que essas espécies possuem um ancestral comum em cujo genoma o evento de inserção ocorreu. Nenhum análogo de evolução convergente altera este quadro, pois não há pressão seletiva para *sítio* de inserção.
- A inserção de SINES parece ser irreversível: nenhum mecanismo de *perda* de SINES é conhecido, a não ser eventos raros de deleções em larga escala de porções do DNA que incluem o SINE. Dessa forma, se duas espécies possuem um SINE em um *locus* comum, a *ausência* deste SINE em uma terceira espécie significa que as duas primeiras estão mais próximas uma da outra do que da terceira.
- Os SINES não apenas indicam parentescos, mas também indicam qual espécie surgiu primeiro. O último ancestral de espécies contendo um SINE comum deve ter surgido *depois* do último ancestral comum que liga essas espécies e outra que não contém este SINE.

N. Okada e colaboradores aplicaram as seqüências SINES a problemas de filogenia.

As baleias, como os australianos, são animais mamíferos que adotaram um estilo de vida aquático. Mas – no caso das baleias – quais são seus parentes mais próximos que ainda têm hábitos terrestres? A paleontologia clássica relaciona a ordem *Cetacea* – compreendendo baleias, golfinhos e toninhas* – com a ordem *Artiodactyla*

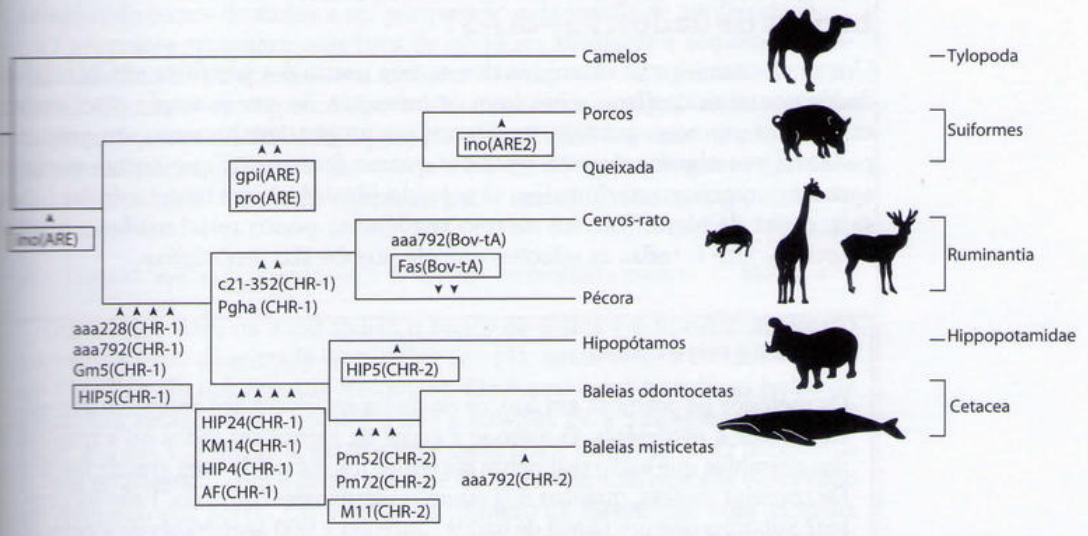


Figura 1.5 Relações filogenéticas entre cetáceos e outros subgrupos dos artiodáctilos, derivadas de análises de seqüências SINES. As pequenas setas indicam eventos de inserção. Cada seta indica a presença de um SINE ou LINE particular, em um *locus* específico em todas as espécies *à direita* das setas. Letras minúsculas identificam os *loci* e letras maiúsculas identificam padrões de seqüências. Por exemplo, o padrão ARE2 aparece apenas em porcos, no *locus* ino. O padrão ARE aparece duas vezes no genoma dos porcos, nos *loci* gpi e pro, e no genoma da queixada, nos mesmos loci. As inserções ARE ocorreram em uma espécie ancestral aos porcos e queixadas, e em nenhuma outra espécie no diagrama. Isso significa que porcos e queixadas são mais próximos um do outro evolutivamente do que de quaisquer dos outros animais estudados. (De Nikaido, M., Rooney, A. P. & Okada, N. (1999), Phylogenetics relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales, *Proceedings of the National Academy of Sciences USA*, **96**, 10261-10266. (Copyright 1999, National Academy of Sciences, USA. Reproduzida com permissão.)

* N. de T. As toninhas se diferem do golfinho pelo focinho curto.

– mamíferos ungulados com um número par de dedos nas patas (incluindo os bovinos). Os cetáceos teriam divergido antes do ancestral comum das três subordens artiodáctilas existentes: *Suiformes* (porcos), *Tylopoda* (incluindo camelos e lhamas) e *Ruminantia* (incluindo cervos, gado, bodes, ovelhas, antílopes, girafas, etc.). Para acomodar os cetáceos de forma correta entre esses grupos, foram realizados vários estudos com seqüências de DNA. Comparações de DNA mitocondrial, dos genes da ribonuclease pancreática, do γ-fibrinogênio e de outras proteínas sugeriram que os parentes mais próximos das baleias são os hipopótamos, e que os cetáceos e hipopótamos formam um grupo separado dentro dos artiodáctilos, estando mais próximos aos *Ruminantia* (ver *Weblema** 1.7).

A análise de SINES confirma este parentesco. Diversos SINES são comuns aos *Ruminantia*, hipopótamos e cetáceos. Quatro SINES aparecem somente nos hipopótamos e cetáceos. Estas observações resultam na árvore filogenética mostrada na Figura 1.5, onde os eventos de inserções de SINES estão indicados.

Fósseis recém-descobertos de ancestrais de hábitos terrestres das baleias confirmam a ligação entre baleias e artiodáctilos. Este é um bom exemplo da complementaridade entre métodos moleculares e paleontológicos: a análise de seqüências de DNA pode especificar as relações entre espécies vivas de forma bastante precisa, mas os fósseis revelam as relações com os seus ancestrais *extintos*.

Pesquisa por seqüências similares em bancos de dados: PSI-BLAST

Um tema comum nos exemplos que temos usado é a pesquisa em bancos de dados por itens similares a um item de interesse. Se, por exemplo, você estiver estudando um novo genoma, ou identificar no genoma humano um gene responsável por alguma doença, você vai querer determinar que outras espécies apresentam genes relacionados. O método ideal deve ser tanto **sensível** – ou seja, capaz de identificar até mesmo seqüências pouco relacionadas – quanto **seletivo** – isto é, todas as relações que ele propõe são verdadeiras.

Precisão e seletividade

Os métodos de pesquisa em bancos de dados envolvem um compromisso entre *precisão* e *seletividade*. O método é capaz de identificar todos ou a maioria dos exemplos que estão realmente presentes, ou ele perde uma grande fração? De maneira inversa, quantos dos *exemplos identificados* (ou "hits") são incorretos? Suponha que um banco de dados contenha 1.000 seqüências de globinas. Suponha também que uma pesquisa por globinas neste banco de dados liste 900 resultados, 700 dos quais são realmente globinas e 200 não. A conclusão seria que este resultado apresenta 300 falso-negativos (seqüências não-identificadas) e 200 falso-positivos. Diminuindo o limiar de tolerância, aumentará o número de *ambos*, os falso-negativos e os falso-positivos. Muitas vezes, é desejável trabalhar com limiares baixos para garantir que nada de importante seja perdido; porém, isso exige uma examinação detalhada dos resultados a fim de se eliminar os falso-positivos.

* N. de T. A palavra em inglês "Weblem", uma combinação das palavras "Web" e "Problem", está sendo traduzida como Weblema em português e significa um problema que necessita dos recursos da *web* para ser resolvido.

Uma ferramenta bastante robusta para pesquisar seqüências em bancos de dados a partir de uma seqüência sonda é o PSI-BLAST (Position Specific Iterated-Basic Local Alignment Search Tool) do NCBI. Um programa anterior, BLAST, funcionava identificando regiões de similaridade locais, sem lacunas, para então reuni-las. O PSI no nome PSI-BLAST refere-se a melhorias que identificam padrões nas seqüências, nos estágios iniciais da pesquisa no banco de dados, para depois refiná-la progressivamente. O reconhecimento de padrões conservados pode aumentar tanto a seletividade quanto a precisão da pesquisa. O PSI-BLAST envolve um processo repetitivo, ou iterativo, como diz o próprio nome, já que os padrões descobertos se tornam cada vez mais definidos em estágios sucessivos da pesquisa.

Estudo de Caso 1.4: Homólogos do gene PAX-6 humano

Os genes PAX-6 controlam o desenvolvimento dos olhos em um conjunto de espécies bastante divergentes (ver Quadro). O gene PAX-6 humano codifica a proteína depositada no SWISS-PROT sob o número de entrada P26367. Para executar o PSI-BLAST, vá para a seguinte URL: `http://www.ncbi.nlm.nih.gov/blast/index.shtml` e selecione PHI- e PSI-BLAST, em "Protein". Cole a seqüência no campo "Search" e use as opções-padrão para as seleções do banco de dados a ser pesquisado e da matriz de similaridade.

O programa retornará uma lista de entradas similares à seqüência sonda, classificadas de acordo com a ordem decrescente da significância estatística (Extratos da resposta obtida são mostrados no Quadro Resultados da pesquisa com o PSI-BLAST pela proteína PAX-6 humana na página 55. Apenas algumas linhas são mostradas, meramente para ilustrar o formato dos resultados.) Uma linha típica, da parte mais inferior na lista, tem este formato:

```
pir||I45557 eyeless, long form - fruit fly (Drosophila melano... 255 7e-67
```

O primeiro item da linha indica o banco de dados e o número de acesso correspondente da entrada (separados por ||), neste caso, a entrada I45557 do PIR (Protein Information Resource). Ela é a proteína homóloga *eyeless* de *Drosophila melanogaster*. O número 255 é o escore para o alinhamento detectado, e a sua significância estatística é dada por $E = 7\times 10^{-67}$. O parâmetro E está relacionado com a probabilidade de o grau de similaridade observado ter ocorrido por acaso. Ou seja, E é o número de seqüências com as quais se esperaria um alinhamento igual ou melhor do que aquele obtido com a seqüência sonda, se o mesmo banco de dados fosse pesquisado com seqüências aleatórias. $E = 7 \times 10^{-67}$ significa que é *extremamente* improvável que mesmo *uma* seqüência aleatória se alinharia tão bem quanto a homóloga de *Drosophila*. Valores de E abaixo de 0,05 seriam considerados significativos; eles deveriam ao menos ser considerados. Para casos duvidosos, você poderia perguntar: Os malpareamentos são conservados? Existe algum padrão ou os pareamentos e malpareamentos estão distribuídos aleatoriamente ao longo das seqüências? Existe um conceito vago, difícil de compreender, denominado *textura* ou perfil do alinhamento, com o qual você irá se familiarizar. A solução definitiva reside em saber se as estruturas são similares, mas esta informação nem sempre está disponível.

→

Note que, se existirem muitas seqüências no banco de dados que sejam bastante similares à seqüência sonda, elas estarão no topo da lista de resultados. Neste exemplo, existem muitos genes PAX similares em outros vertebrados. Você terá de examinar cuidadosamente até o final da lista para encontrar parentes distantes que considera relevantes.

Mesmo no caso da proteína *eyeless* de *Drosophila*, um parente bastante próximo da seqüência sonda, o programa relata apenas um alinhamento local para uma parte das seqüências. O alinhamento completo é mostrado no Quadro Alinhamento de seqüência par a par completo das proteínas PAX-6 humana e *eyeless* de *Drosophila melanogaster*, na página 57.

Et in terra PAX hominibus, muscisque...

Os olhos dos humanos, da mosca e do polvo são estruturalmente bastante diferentes. O conhecimento convencional, observando a imensa vantagem seletiva conferida pela habilidade de enxergar, afirmava que os olhos se desenvolveram de forma independente em diversos filos. Foi, portanto, uma grande surpresa quando se descobriu que um gene responsável pelo controle do desenvolvimento do olho em humanos possui um homólogo responsável pelo desenvolvimento do olho em *Drosophila*.

Os genes PAX-6 humano e de camundongo foram os primeiros a serem clonados. PAX-6 é um gene regulador do desenvolvimento, que controla uma cascata complexa de eventos no desenvolvimento do olho. Mutações no gene humano causam uma condição clínica chamada *aniridia*, um defeito no desenvolvimento do olho em que a íris encontra-se ausente ou deformada. O homólogo PAX-6 em *Drosophila* – chamado de gene *eyeless* – possui a função similar de controlar o desenvolvimento ocular. Moscas com mutações nesse gene não desenvolvem olhos; de forma semelhante, a expressão desse gene nas asas, pernas ou antenas da mosca produz olhos ectópicos (fora do lugar). (O mutante *eyeless* de *Drosophila* foi descrito pela primeira vez em 1915. Pouco se suspeitou na época da relação com genes de mamíferos.)

Os genes de insetos e mamíferos não apenas têm seqüências similares, mas estão relacionados em tal grau que as suas funções cruzam barreiras interespecíficas. A expressão do gene PAX-6 de camundongo em *Drosophila* leva ao desenvolvimento de olhos ectópicos, da mesma forma que a expressão do gene *eyeless* da própria mosca.

PAX-6 possui homólogos em outros filos, incluindo vermes achatados, ascídias, ouriços-do-mar e nematódeos. A observação de que as rodopsinas – uma família de proteínas que contêm retinal como cromóforo – funcionam como pigmentos sensíveis à luz em diferentes filos é uma evidência de suporte para uma origem comum dos diferentes sistemas fotorreceptores. As diferenças estruturais genuínas na anatomia macroscópica dos diferentes olhos refletem a divergência e o desenvolvimento independente de estrutura de ordem superior.

Resultados da pesquisa com o PSI-BLAST pela proteína PAX-6 humana

Cinco iterações do PSI-BLAST foram realizadas utilizando a PAX-6 humana como seqüência sonda na pesquisa ao banco de dados não-redundante (nr). O banco de dados nr do NCBI é um conjunto de seqüências únicas, selecionadas a partir dos bancos de dados completos, com a finalidade de eliminar redundâncias. O arquivo de saída contém uma lista das seqüências similares identificadas no banco de dados. Ele também contém os alinhamentos par a par das regiões bem alinhadas entre a seqüência sonda e as seqüências encontradas. Três alinhamentos selecionados são mostrados aqui: PAX-6 de *Danio rerio*, $E = 10^{-134}$; *eyeless* de *Drosophila*, $E = 7 \times 10^{-67}$; e outra proteína de *Drosophila*, *even skipped*, com $E = 0{,}001$, cujo alinhamento é mais curto e menos perfeito:

Escore = 46,9 bits (110) e Valor E = 0,001

```
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.
RID: 1081182496-14481-96333547442.BLASTQ3
Query= sp|P26367|PAX6_HUMAN Paired box protein Pax-6(Oculorhombin)
(Aniridia, type II protein) - Homo sapiens (Human).
        (422 letters)

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
          2,738,511 sequences; 768,166,133 total letters

Results of PSI-Blast iteration 5
Sequences with E-value BETTER than threshold
Sequences producing significant alignments:                    (bits) Value

gi|2696972|dbj|BAA24025.1| PAX6 SL [Cynops pyrrhogaster]          499   e-140
gi|189353|gb|AAA59962.1| oculorhombin                            496   e-139
gi|4505615|ref|NP_000271.1| paired box gene 6 isoform a; Paired box h...  495   e-139
gi|27469846|gb|AAH41712.1| MGC52531 protein [Xenopus laevis]     495   e-139
gi|6981334|ref|NP_037133.1| paired box gene 6; paired box homeotic ge...  494   e-138
gi|2696971|dbj|BAA24024.1| PAX6 LL [Cynops pyrrhogaster]         491   e-137
gi|7305369|ref|NP_038655.1| paired box gene 6; small eye; Dickie's sm...  491   e-137
gi|42592306|emb|CAF29075.1| putative pax6 isoform 5a [Rattus norvegicus]  491   e-137
gi|1685047|gb|AAB36681.1| paired-type homeodomain Pax-6 protein [Xeno...  491   e-137
gi|26389393|dbj|BAC25729.1| unnamed protein product [Mus musculus]  490   e-137
gi|4580424|ref|NP_001595.2| paired box gene 6 isoform b; Paired box h...  490   e-137
gi|8132383|gb|AAF73271.1| paired domain transcription factor variant ...  490   e-137
gi|5758939|gb|AAD50903.1| paired-box transcription factor -+ isoform ...  489   e-137
gi|18138028|emb|CAC80516.1| paired box protein [Mus musculus]    489   e-137
gi|44922124|gb|AAS48919.1| paired box 6 isoform 5a [Rattus norvegicus]  489   e-137
gi|45384210|ref|NP_990397.1| PAX6 protein [Gallus gallus]        489   e-137
gi|383296|prf||1902328A PAX6 gene                                488   e-137
gi|1488322|gb|AAB05932.1| Xpax6 [Xenopus laevis]                 488   e-136
gi|34364871|emb|CAE45868.1| hypothetical protein [Homo sapiens]  488   e-136
gi|2369655|emb|CAA68838.1| PAX-6 protein [Astyanax mexicanus]    487   e-136
```

... resultados adicionais removidos...

... seguem três alinhamentos selecionados...

Resultados da pesquisa com o PSI-BLAST pela proteína PAX-6 humana (*continuação*)

```
                                Alignments
>gi|18859209|ref|NP_571379.1| paired box gene 6a; paired box homeotic
gene 6 [Danio rerio] Length = 451

 Score =  478 bits (1231), Expect = e-134
 Identities = 404/436 (92%), Positives = 409/436 (93%), Gaps = 18/436 (4%)

Query: 1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ------------ 47
           MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ
Sbjct: 20  MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNE 79

Query: 48  -VSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRL 106
            VSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVV KIAQYKRECPSIFAWEIRDRL
Sbjct: 80  NVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVGKIAQYKRECPSIFAWEIRDRL 139

Query: 107 LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGT 166
           LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMY+KLRMLNGQTG+WGTRPGWYPGT
Sbjct: 140 LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYEKLRMLNGQTGTWGTRPGWYPGT 199

Query: 167 SVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE 226
           SVPGQP QDGCQQ +GGGENTNSISSNGEDSDE QMRLQLKRKLQRNRTSFTQEQIEALE
Sbjct: 200 SVPGQPNQDGCQQSDGGGENTNSISSNGEDSDETQMRLQLKRKLQRNRTSFTQEQIEALE 259

Query: 227 KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIP 286
           KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN+ SHIP
Sbjct: 260 KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNSSSHIP 319

Query: 287 ISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPP 346
           ISSSFSTSVYQPIPQPTTPV SFTSGSMLGR+DTALTNTYSALPPMPSFTMANNLPMQP
Sbjct: 320 ISSSFSTSVYQPIPQPTTPV-SFTSGSMLGRSDTALTNTYSALPPMPSFTMANNLPMQP- 377

Query: 347 VPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQ 406
             SQTSSYSCMLPTSPSVNGRSYDTYTPPHMQ HMNSQ M  SGTTSTGLISPGVSVPVQ
Sbjct: 378 --SQTSSYSCMLPTSPSVNGRSYDTYTPPHMQAHMNSQSMAASGTTSTGLISPGVSVPVQ 435

Query: 407 VPGSEPDMSQYWPRLQ 422
           VPGSEPDMSQYWPRLQ
Sbjct: 436 VPGSEPDMSQYWPRLQ 451


>pir||I45557 eyeless, long form - fruit fly (Drosophila melanogaster)
 emb|CAA56038.1| (X79493) transcription factor [Drosophila melanogaster]
          Length = 838

 Score =  255 bits (644), Expect = 7e-67
 Identities = 124/132 (93%), Positives = 128/132 (96%)

Query: 5   HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64
           HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 38  HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 97

Query: 65  SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN 124
           SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNIPSVSSIN
Sbjct: 98  SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSIN 157

Query: 125 RVLRNLASEKQQ 136
           RVLRNLA++K+Q
Sbjct: 158 RVLRNLAAQKEQ 169
```

```
>gi|24652276|ref|NP_523670.2| CG2328-PA [Drosophila melanogaster]
 gi|12644102|sp|P06602|HMEV_DROME Segmentation protein even-skipped
 gi|7958|emb|CAA28784.1| even skipped [Drosophila melanogaster]
 gi|7303818|gb|AAF58865.1| CG2328-PA [Drosophila melanogaster]
          Length = 376

 Score = 46.9 bits (110), Expect = 0.001
 Identities = 33/95 (34%), Positives = 51/95 (53%)

Query: 174 QDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTH 233
           Q G  Q      N    S +   +      +     ++R RT+FT++Q+  LEKEF + +
Sbjct: 34  QYGKPQTPPPSPNECLSSPDNSLNGSRGSEIPADPSVRRYRTAFTRDQLGRLEKEFYKEN 93

Query: 234 YPDVFARERLAAKIDLPEARIQVWFSNRRAKWRRE 268
           Y     R  LAA+++LPE+ I+VWF NRR K +R+
Sbjct: 94  YVSRPRRCELAAQLNLPESTIKVWFQNRRMKDKRQ 128
```

Alinhamento de seqüência par a par completo das proteínas PAX-6 humana e *eyeless* de *Drosophila melanogaster*

```
PAX6_human      ---------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTRQ 27
eyeless         MFTLQPTPTAIGTVVPPWSAGTLIERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTRQ 60
                                                ::.************.**********

PAX6_human      KIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKI 87
eyeless         KIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKI 120
                ************************************************.******

PAX6_human      AQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQ----------- 136
eyeless         SQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTS 180
                :******************.*.*********************::*:*

PAX6_human      ------------MG-------------------------------------------ADG 141
eyeless         AGNSISAKVSVSIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGATNSGEGSEQEA 240
                           :*                                           :.

PAX6_human      MYDKLRMLNGQTGS--------------------WGTRP---------------------- 160
eyeless         IYEKLRLLNTQHAAGPGPLEPARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSW 300
                :*:***:** * .:                    ***.

PAX6_human      -------GWYPG-------TSVP-----------------------------GQP---- 172
eyeless         PPRHYSGSWYPTSLSEIPISSAPNIASVTAYASGPSLAHSLSPPNDIKSLASIGHQRNCP 360
                       .***        :*.*                                 *:

PAX6_human      ----------TQDGCQQQEGG---GENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQ 219
eyeless         VATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDDQARLILKRKLQRNRTSFTN 420
                          ** *.:* *   ***:*. :**  :::: * ** *************:

PAX6_human      EQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQAS 279
eyeless         DQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEKLRNQRRTPN 480
                :**::*********************.**.*************************** ..

PAX6_human      NTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLG---------------------- 316
eyeless         STGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPT 540
                .* :  . **: :*:     *:. :. **: ***  *

PAX6_human      ------------------------------------------------------------
eyeless         LGAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHA 600

PAX6_human      ----------------------------RTDTALTNTYSALPPMPSFTMANNLPMQPPVP 348
eyeless         QGHALVPAISPRLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSP 660
                                          :*  :::::*.*:.*:***. :   *: ** *

PAX6_human      S-------QTSSYSCMLPTSP-------------------------------SVNGRS 368
eyeless         IPQQGDLTPSSLYPCHMTLRPPPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSG 720
                      :* *.* :.  *                                  * .* .

PAX6_human      YDTYTP----------------------------PHMQTHMNSQP----------MGTS 389
eyeless         YEVLSAYALPPPPMASSSAADSSFSAASSASANVTPHHTIAQESCPSPCSSASHFGVAHS 780
                *:. :.                              **     :* *          :. *

PAX6_human      GTTSTGLISPGVS----------------VPVQVPGS----EPDMSQYWPRLQ----- 422
eyeless         SGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHVAPGKQQFFASCFYSPWV 838
                .  *:. ***.**                .* ...*:     *. .*::.
```

Espécies reconhecidas pelo PSI-BLAST ("hits") como similares à seqüência sonda PAX-6 humana

Achaearanea tepidariorum
Acipenser baerii
Acropora millepora
Acropora palmata
Ambystoma mexicanum
Anopheles gambiae
Archaster typicus
Archegozetes longisetosus
Artemia franciscana
Astatoreochromis alluaudi
Astatotilapia brownae
Asterina miniata
Astyanax mexicanus
Bos taurus
Branchiostoma belcheri
Branchiostoma floridae
Branchiostoma lanceolatum
Caenorhabditis briggsae
Caenorhabditis elegans
Canis familiaris
Carassius auratus
Cassiopea xamachana
Chrysaora quinquecirrha
Ciona intestinalis
Ciona savignyi
Cladonema californicum
Coturnix coturnix
Crocodylus niloticus
Ctenophorus ornatus
Cupiennius salei
Cynops pyrrhogaster
Cyprinus carpio
Danio rerio
Discocelis tigrina
Drosophila ananassae
Drosophila mauritiana
Drosophila melanogaster
Drosophila sechellia
Drosophila simulans
Drosophila virilis
Dugesia japonica
Echinococcus granulosus
Eleutheria dichotoma
Eleutherodactylus coqui
Emys orbicularis
Ephydatia fluviatilis
Ephydatia muelleri
Equus caballus
Eremothecium gossypii
Euprymna scolopes
Gallotia stehlini
Gallus gallus
Gasterosteus aculeatus
Girardia tigrina
Gryllus bimaculatus
Haliotis asinina
Haliotis rufescens
Halocynthia roretzi
Heliocidaris erythrogramma
Heliocidaris tuberculata
Helobdella robusta
Helobdella triserialis
Hemicentrotus pulcherrimus
Herdmania curvata
Heterodontus francisci
Homo sapiens
Hydra littoralis
Hydra magnipapillata
Hydra vulgaris
Hydractinia symbiolongicarpus
Ilyanassa obsoleta
Labidochromis caeruleus
Lampetra fluviatilis
Lampetra planeri
Latimeria menadoensis
Lethenteron japonicum
Leucopsarion petersii
Lineus sanguineus
Lingula anatina
Loligo opalescens
Lytechinus variegatus
Mesocricetus auratus
Monodelphis domestica
Mus musculus
Nematostella vectensis
Nereis virens
Oreochromis aureus
Oreochromis niloticus
Oryctolagus cuniculus
Oryzias latipes
Paracentrotus lividus
Paralichthys olivaceus
Patella vulgata
Petromyzon marinus
Phallusia mammilata
Phascolion strombi
Platynereis dumerilii
Pleurodeles waltl
Podocoryne carnea
Pristina leidyi
Ptychodera flava
Raja eglanteria
Rattus norvegicus
Related structures
Saccoglossus kowalevskii
Schistocerca americana
Scyliorhinus canicula
Spongilla lacustris
Stichopus japonicus
Strongylocentrotus purpuratus
Sus scrofa
Takifugu rubripes
Tetranychus urticae
Theromyzon trizonare
Tribolium castaneum
Trichoplax adhaerens
Tripedalia cystophora
Triturus alpestris
Tropheus duboisi
Xenopus laevis
Xenopus tropicalis
Xiphophorus hellerii

Exemplo PERL 1.3 Quais espécies detectadas pelo PSI-BLAST contêm homólogas da PAX-6 humana?

O PSI-BLAST relata as espécies em que foram identificadas seqüências (ver Quadro Resultados da pesquisa com o PSI-BLAST pela proteína PAX-6 humana). Estas informações aparecem entre colchetes no texto do arquivo de saída; por exemplo:

```
emb|CAA56038.1| (X79493) transcription factor [Drosophila melanogaster]
```

(Na seção contendo os valores de *E*, os nomes das espécies podem estar truncados.)

→

O programa PERL a seguir extrai os nomes das espécies do arquivo de saída do PSI-BLAST.

```
#!/usr/bin/perl
#extract species from psiblast output

#Method:
#  For each line of input, check for a pattern of form [Drosophila melanogaster]
#  Use each pattern found as the index in an associative array
#  The value corresponding to this index is irrelevant
#  By using an associative array, subsequent instances of the same
#     species will overwrite the first instance, keeping only a unique set
#  After processing of input complete, sort results and print.


while (<>) {                          # read line of input
    if (/\[([A-Z][a-z]+ [a-z]+)\]/) { # select lines containing strings of form
                                      #     [Drosophila melanogaster]
     $species{$1} = 1;                # make or overwrite entry in
   }                                  #         associative array
}

foreach (sort(keys(%species))){       # in alphabetical order,
    print "$_\n";                     #    print species names
}
```

Foram encontradas 122 espécies (ver Quadro: Espécies reconhecidas pelo PSI-BLAST ("hits") como similares à seqüência sonda PAX-6 humana).

O programa utiliza os excelentes recursos de reconhecimento de padrões do PERL para localizar seqüências de caracteres da forma `[Drosophila melanogaster]`. Queremos especificar o seguinte padrão:

- um colchete para abrir,
- seguido de uma palavra iniciando com letra maiúscula, seguida por um número variável de letras minúsculas,
- então, um espaço entre palavras,
- depois, uma palavra inteira em letras minúsculas,
- e, finalmente, um colchete para fechar.

Este tipo de padrão é denominado **expressão regular** e aparece no programa PERL da seguinte forma: `[([A-Z][a-z]+ [a-z]+)]`.

Cada um dos blocos da expressão acima especifica um grupo de caracteres:

`[A-Z]` = qualquer letra na faixa A, B, C, ... Z

`[a-z]` = qualquer letra na faixa a, b, c, ... z

Podemos especificar repetições:

`[A-Z]` = *uma* letra maiúscula

`[a-z]+` = *uma ou mais* letras minúsculas

e combinar os resultados:

`[A-Z][a-z]+ [a-z]+` = uma letra maiúscula seguida de uma ou mais letras minúsculas (o nome do gênero), seguida de um espaço, seguido de uma ou mais letras minúsculas (o nome da espécie).

Exemplo PERL 1.3 (*continuação*)

Colocar a expressão entre parênteses: `([A-Z][a-z]+ [a-z]+)` diz ao PERL para salvar os dados que conferem com o padrão para consultas futuras. No programa PERL, esses dados que conferem com o padrão são designados pela variável `$1`. Assim, se a linha de entrada contiver `[Drosophila melanogaster]`, a instrução

```
$species{$1} = 1;
```

resultará efetivamente em:

```
$species{"Drosophila melanogaster"} = 1;
```

Por fim, queremos incluir os colchetes envolvendo os nomes do gênero e da espécie, mas colchetes significam faixas de caracteres. É necessário, portanto, colocar barras antes dos colchetes: `\[...\]` para obter o padrão final: `\[([A-Z][a-z]+ [a-z]+)\]`.

O uso da matriz associativa para preservar somente um único conjunto de espécies é outro aspecto instrutivo do programa. Lembre-se de que uma matriz associativa é uma generalização de uma matriz comum ou vetor, em que os elementos não são indexados por números inteiros, mas por seqüências de caracteres arbitrárias. Uma segunda consulta a uma matriz associativa, com uma seqüência de índices previamente encontrada, talvez pudesse alterar o valor na matriz, mas não a lista de seqüências de índices. Neste caso, não nos preocupamos com o valor, apenas utilizamos as seqüências de índices para compilar uma lista única de espécies detectadas. Referências múltiplas à mesma espécie irão apenas sobrescrever a primeira referência, e *não* originar uma lista repetitiva.

Versões mais atuais do PSI-BLAST descrevem também a distribuição taxonômica das espécies encontradas nas pesquisas. Contudo, o programa PERL fornecido neste exemplo poderia ser útil caso se queira obter os alinhamentos, ou realizar outros tipos de análises dos resultados.

Introdução à estrutura de proteínas

Com a estrutura de proteínas, deixamos para trás o mundo unidimensional de seqüências de nucleotídeos e aminoácidos e entramos no mundo tridimensional das estruturas moleculares. Alguns dos recursos para armazenamento e recuperação de informações biológicas moleculares sobrevivem a esta mudança relativamente intactos, alguns são substancialmente alterados e outros desaparecem completamente.

Bioquimicamente, as proteínas desempenham uma variedade de papéis nos processos vitais: há proteínas estruturais (por exemplo, as proteínas do capsídeo viral, a camada mais externa e áspera da pele de humanos e de outros animais, e as proteínas do citoesqueleto); proteínas que catalisam reações químicas (as enzimas); proteínas de transporte e de armazenagem (hemoglobina e ferritina); proteínas reguladoras, incluindo hormônios e proteínas sinalizadoras ou receptoras de sinais; proteínas que controlam a transcrição gênica; e proteínas envolvidas em reconhecimento, incluindo moléculas de adesão celular, anticorpos e outras proteínas do sistema imune.

Proteínas são moléculas grandes. Em muitos casos, apenas uma pequena porção da estrutura – o **sítio ativo** – funciona de maneira precisa, com o restante da estrutura existindo principalmente para criar e manter as relações espaciais entre os resíduos do sítio ativo. As proteínas evoluíram por meio de alterações estrutu-

rais, originadas por mutações na seqüência de aminoácidos e rearranjos gênicos, que integram diferentes combinações de subunidades estruturais.

Atualmente são conhecidas cerca de 30.000 estruturas de proteínas.* A maioria foi determinada por cristalografia por difração de raios X ou por ressonância magnética nuclear (NMR – *nuclear magnetic ressonance*). A partir destas estruturas, obtivemos o nosso conhecimento acerca das funções de proteínas individuais – por exemplo, a explicação química da atividade catalítica das enzimas – e dos princípios gerais de estrutura e enovelamento de proteínas.

Do ponto de vista químico, as moléculas de proteínas são polímeros longos contendo, tipicamente, milhares de átomos, compostos por um **esqueleto** repetitivo uniforme (ou **cadeia principal**) com uma **cadeia lateral** específica ligada a cada resíduo (ver Figura 1.6). A seqüência de aminoácidos de uma proteína equivale a uma série de cadeias laterais.

A cadeia polipeptídica se enovela no espaço; o seu percurso, ou o arranjo da cadeia principal, define um **padrão de enovelamento**. As proteínas apresentam uma grande variedade de padrões de enovelamento. Implícitas a esses padrões há diversas características estruturais comuns. Essas características incluem a recorrência de padrões estruturais explícitos – por exemplo, hélices α e folhas β (Figura 1.7) – e princípios ou características comuns, como o denso empacotamento de átomos no interior das proteínas. O enovelamento pode ser pensado como um tipo de condensação ou cristalização intramolecular (ver Capítulo 5).

A natureza hierárquica da arquitetura de proteínas

O químico de proteínas dinamarquês K. U. Linderstrøm-Lang descreveu os seguintes níveis de estruturas de proteínas: a seqüência de aminoácidos – o conjunto de ligações químicas primárias – é chamada de **estrutura primária**. A atribuição de hélices e folhas – de acordo com o padrão de pontes de hidrogênio da cadeia principal – é denominada **estrutura secundária**. O agrupamento e as interações das hélices e folhas formam a chamada **estrutura terciária**. Para proteínas compostas por mais de uma subunidade, J. D. Bernal denominou o agrupamento dos monômeros **estrutura quaternária**. Em alguns casos, a evolução pode fundir proteínas – transformando a estrutura quaternária em terciária. Por exemplo, cinco enzimas distintas da bactéria *E. coli*, que catalisam etapas sucessivas na via de biossíntese de aminoácidos aromáticos, correspondem a cinco regiões de uma única proteína no fungo *Aspergillus nidulans*. Ocasionalmente, monômeros homólogos formam oligômeros de diferentes maneiras; por exemplo, as globinas formam tetrâmeros em hemoglobinas de mamíferos, e dímeros – utilizando uma interface diferente – no marisco arcado *Scapharca inaequivalvis*.

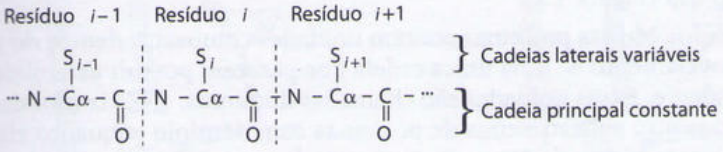


Figura 1.6 As cadeias polipeptídicas de proteínas possuem uma cadeia principal de estrutura constante e cadeias laterais de seqüência variável. Aqui, S_{i-1}, S_i e S_{i+1} representam as cadeias laterais. As cadeias laterais podem ser escolhidas, independentemente, do conjunto de 20 aminoácidos-padrão. É a seqüência de cadeias laterais que confere a cada proteína suas características estruturais e funcionais.

* N. de T. Em agosto de 2007, este número chegava a pouco mais de 41.000 estruturas.

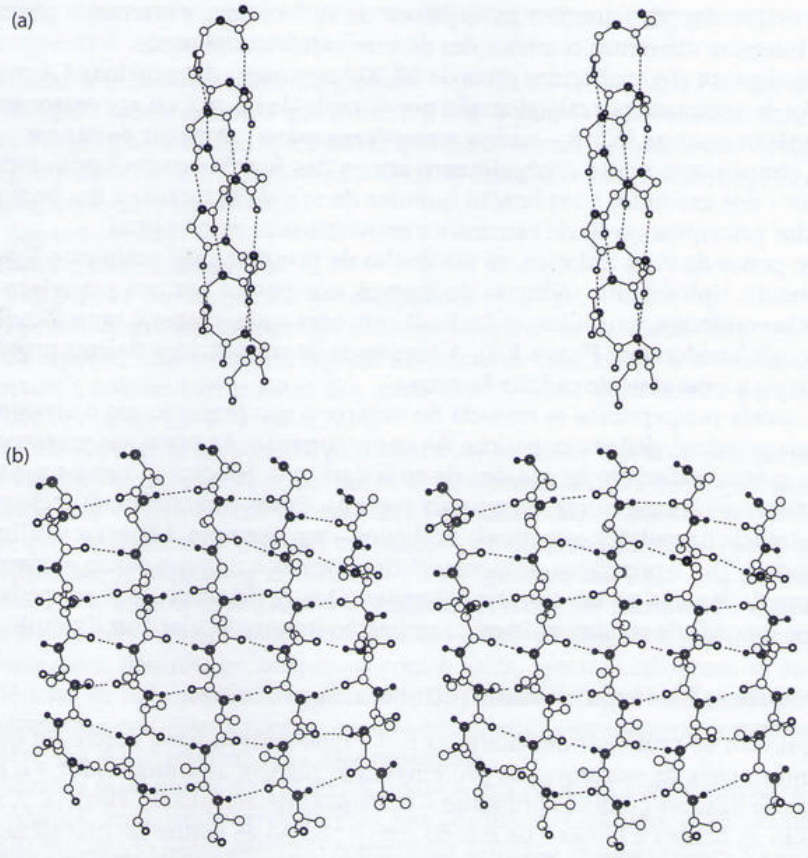


Figura 1.7 Estruturas secundárias regulares de proteínas. (a) Hélice α. (b) Folha β. Átomos de hidrogênio não estão representados. (b) Ilustra uma folha β paralela, nas quais todas as cadeias apontam na mesma direção. Folhas β antiparalelas, nas quais todos os pares de cadeias adjacentes apontam em direções opostas, também são comuns. De fato, as folhas β podem ser formadas por qualquer combinação de cadeias paralelas e antiparalelas.

A utilização de níveis adicionais na hierarquia já se mostrou útil:

- **Estruturas supersecundárias** Proteínas apresentam um padrão comum de interação entre hélices e folhas β que estão próximas na seqüência. Essas *estruturas supersecundárias* incluem o grampo de hélices α, o grampo β e a unidade β-α-β (Figura 1.8).
- **Domínios** Muitas proteínas contêm unidades compactas dentro do padrão de enovelamento de uma única cadeia que parecem possuir estabilidade independente. Essas unidades são chamadas domínios. (Não confunda domínio enquanto subestruturas de proteínas com domínio enquanto classe de organismos vivos: Archaea, Bacteria e Eukarya.) A proteína L1 de ligação ao RNA apresenta características típicas de proteínas multidomínio: o sítio de ligação está situado na fenda criada entre os dois domínios e a geometria relativa dos dois domínios é flexível, permitindo mudanças conformacionais induzidas pelo ligante (Figura 1.9). Na hierarquia, os domínios ficam entre as estruturas supersecundárias e a estrutura terciária de um monômero integral.

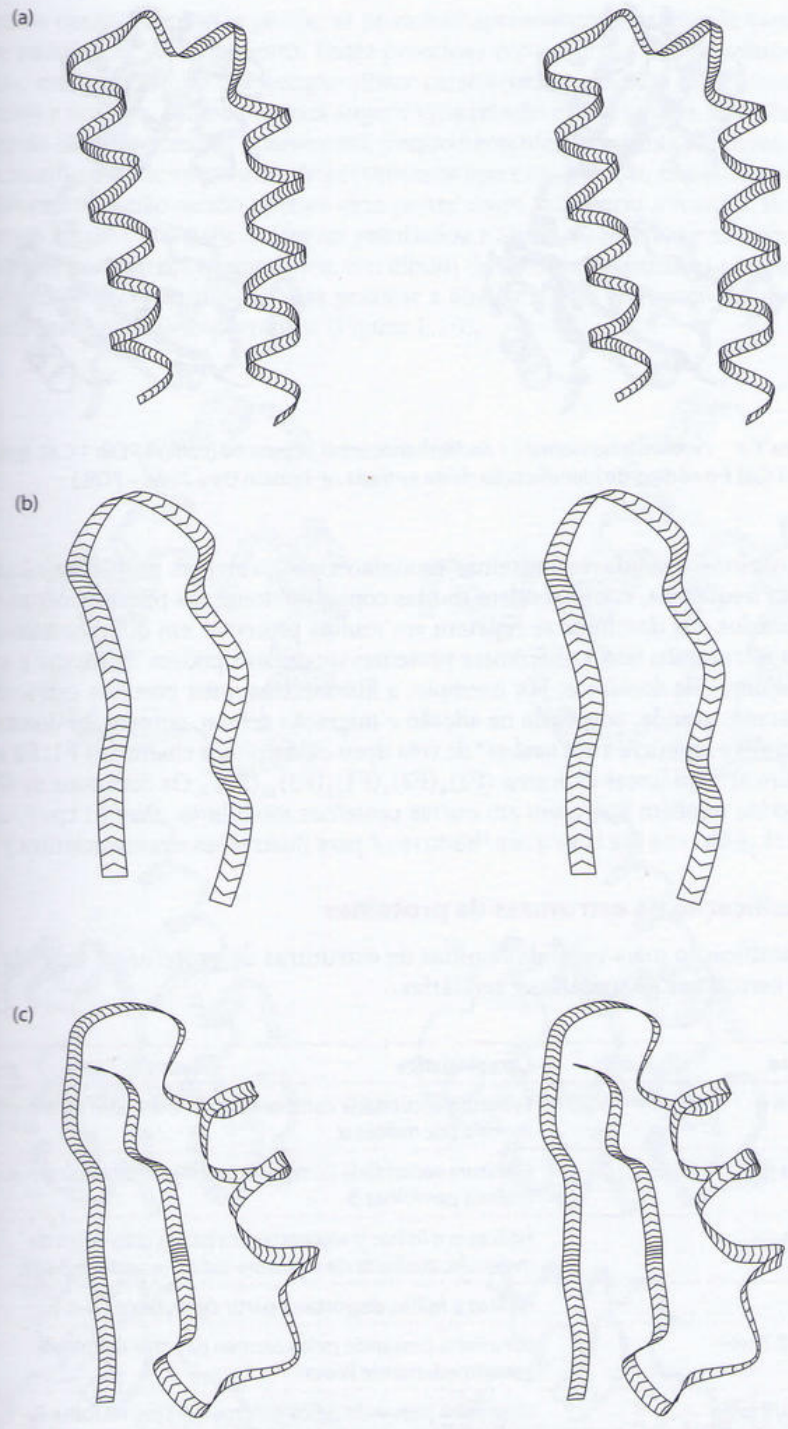


Figura 1.8 Estruturas supersecundárias comuns. (a) Grampo de hélices α (α-helix hairpin), (b) grampo de fitas β (β-hairpin*), (c) unidade β–α–β. As listras anguladas, preenchendo as estruturas, indicam a direção da cadeia.

* N. de T. *Hairpin* significa grampo de cabelo. Este é precisamente o formato das estruturas em (a) e (b) e daí a origem dos seus nomes em português.

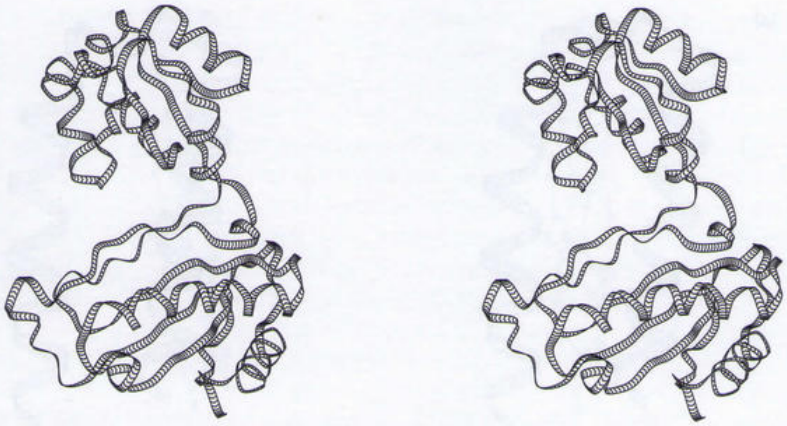


Figura 1.9 Proteína ribossomal L1 de *Methanococcus jannaschii* [código PDB: 1 CJS]. ([código PDB: 1 CJS] é o código de identificação desta entrada no Protein Data Bank – PDB.)

- **Proteínas modulares** Proteínas modulares são proteínas multidomínio que, com freqüência, compreendem muitas cópias de domínios proximamente relacionados. Os domínios se repetem em muitas proteínas em diferentes contextos estruturais; isto é, diferentes proteínas modulares podem "misturar e unir" conjuntos de domínios. Por exemplo, a fibronectina, uma proteína extracelular bastante grande, envolvida na adesão e migração celular, contém 29 domínios, inclusive repetições em *tandem** de três tipos de domínios chamados F1, F2 e F3. É um arranjo linear da forma $(F1)_6(F2)_2(F1)_3(F3)_{15}(F1)_3$. Os domínios da fibronectina também aparecem em outras proteínas modulares. (Ver `http://www.bork.embl-heidelberg.de/Modules/` para ilustrações e nomenclatura.)

Classificação de estruturas de proteínas

A classificação mais geral de famílias de estruturas de proteínas é baseada nas suas estruturas secundárias e terciárias.

Classe	Característica
Hélice α	Estrutura secundária composta exclusivamente ou em sua maioria por hélices α
Folha β	Estrutura secundária composta exclusivamente ou em sua maioria por folhas β
α + β	Hélices α e folhas β separadas em partes diferentes da molécula; ausência da estrutura supersecundária β-α-β
α/β	Hélices e folhas dispostas a partir de unidades β-α-β
α/β linear	Uma linha passando pelos centros das fitas da folha é aproximadamente linear
Barril α/β	Uma linha passando pelos centros das fitas da folha é aproximadamente circular
Pouca ou nenhuma estrutura secundária	

* N. de T. Repetições em *tandem* são repetições dispostas uma atrás da outra.

Dentro dessas categorias gerais, as proteínas apresentam uma grande variedade de padrões de enovelamento. Entre proteínas com padrões de enovelamento similar, existem famílias que compartilham características em suas estruturas, seqüências e funções, suficientes para sugerir uma relação evolucionária. Entretanto, proteínas não relacionadas apresentam, freqüentemente, estruturas similares.

A classificação de estruturas de proteínas ocupa uma posição fundamental na bioinformática, não sendo apenas uma ponte entre seqüência e função. Retornaremos a este tema para descrever resultados e sítios de interesse na *web*. Por enquanto, as ilustrações seguintes, um álbum de algumas estruturas pequenas, fornecem uma oportunidade para praticar a análise visual e o reconhecimento de padrões espaciais importantes (Figura 1.10).

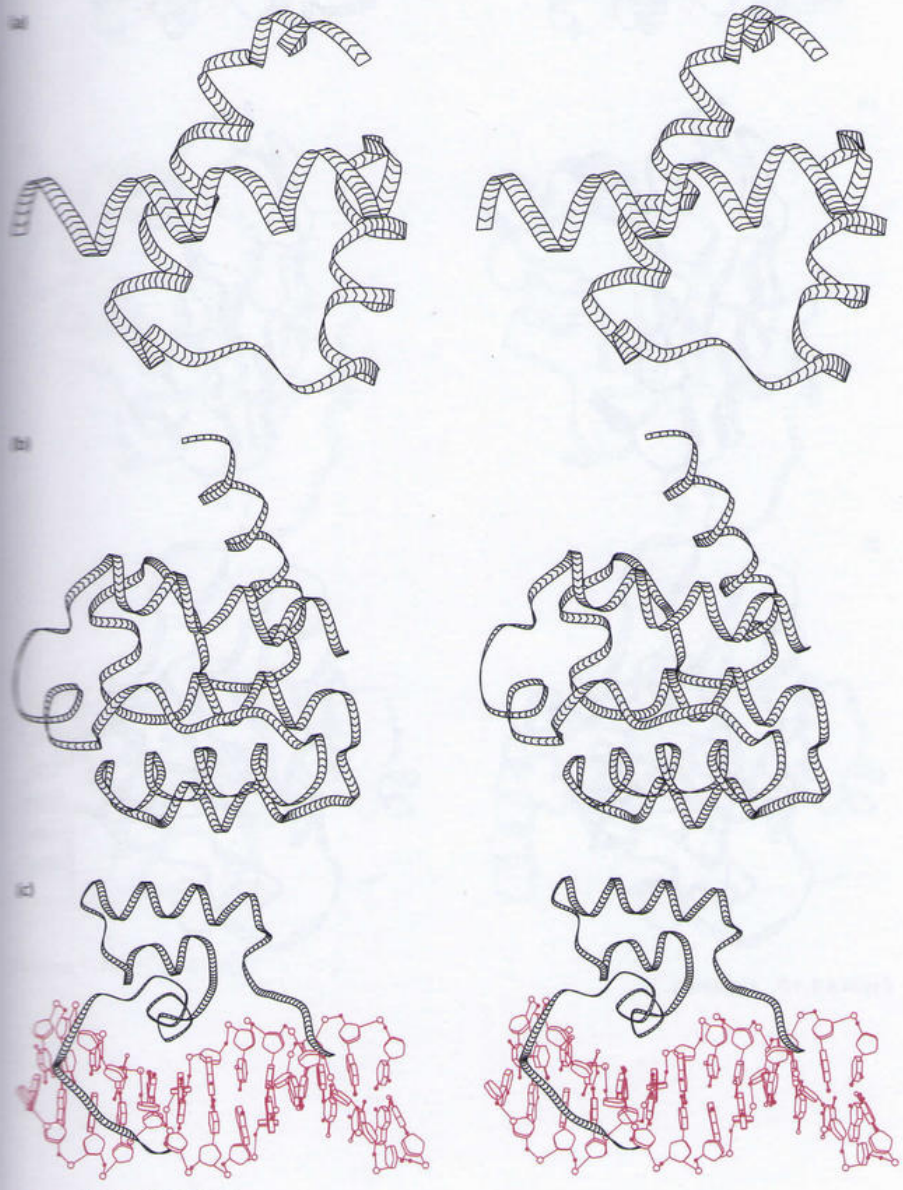


Figura 1.10 *Continua.*

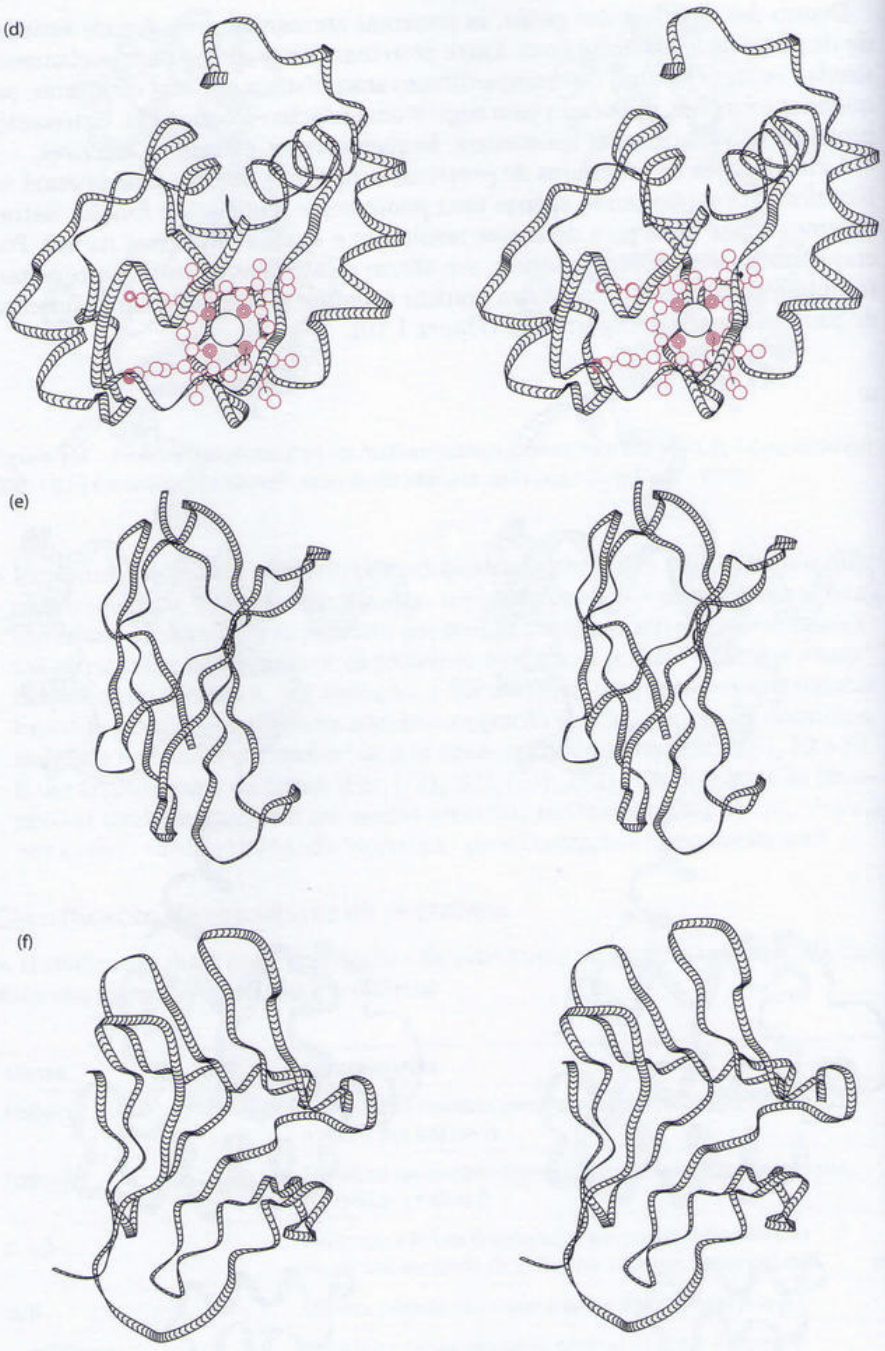


Figura 1.10 *Continua.*

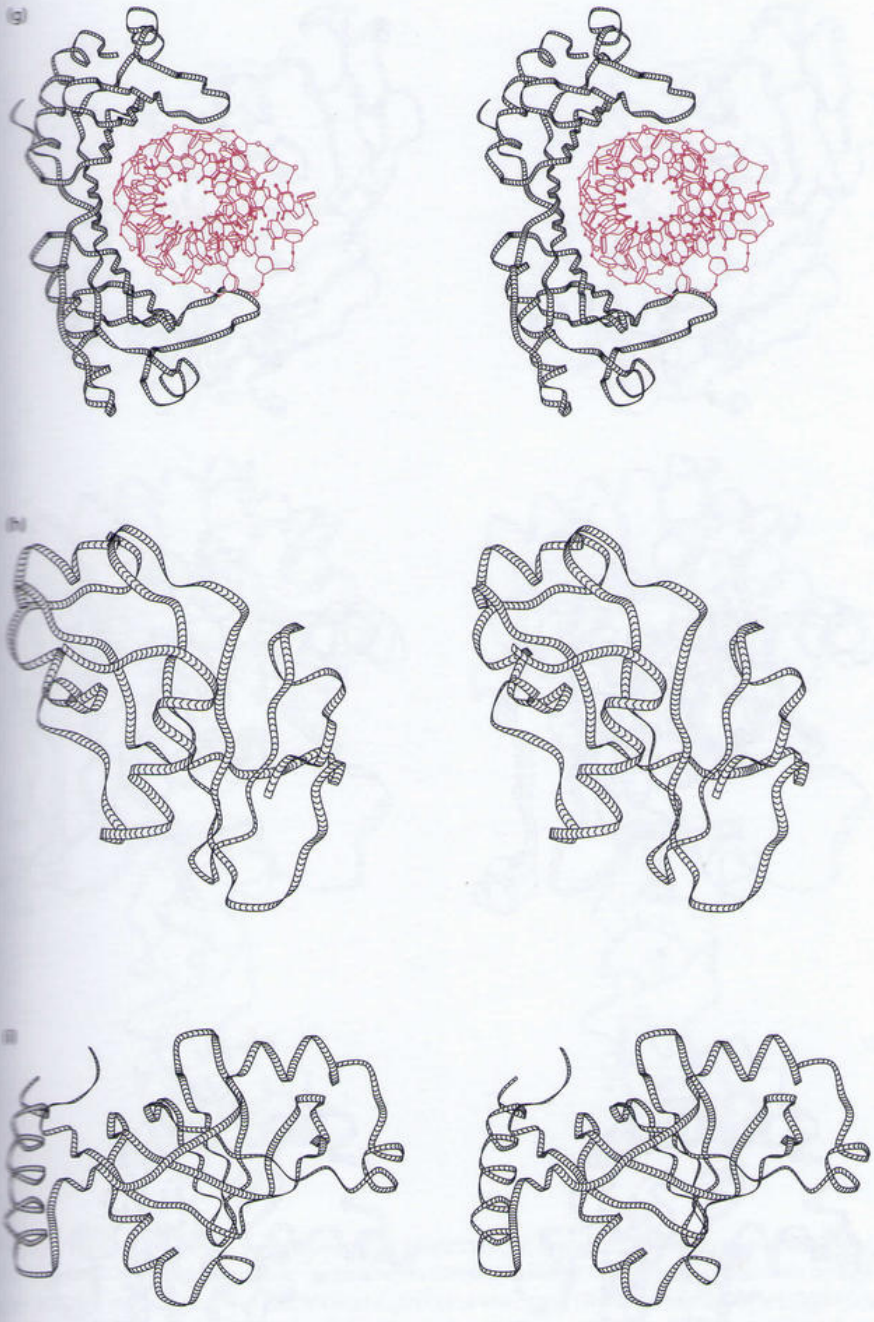


Figura 1.10 *Continua.*

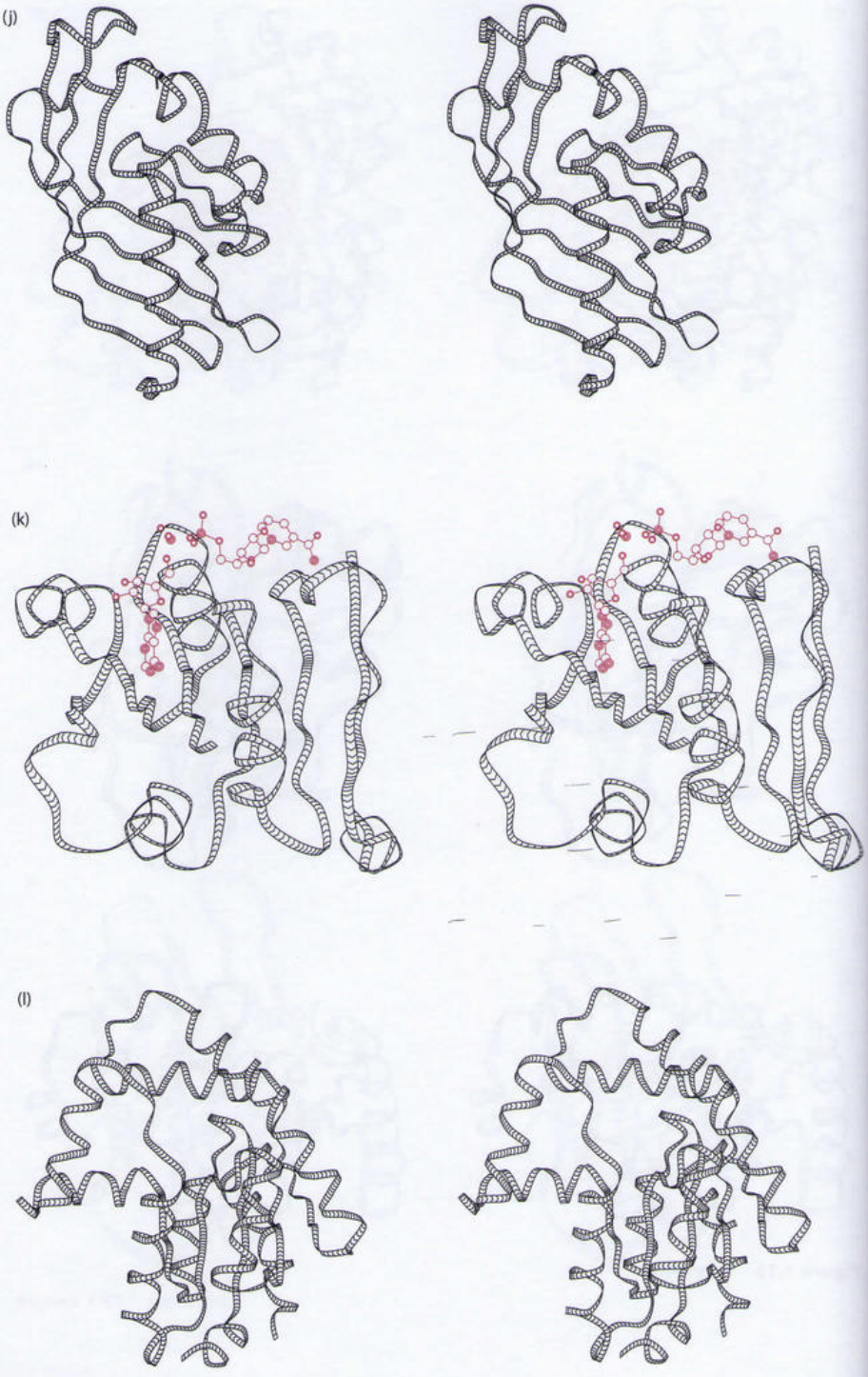


Figura 1.10 *Continua.*

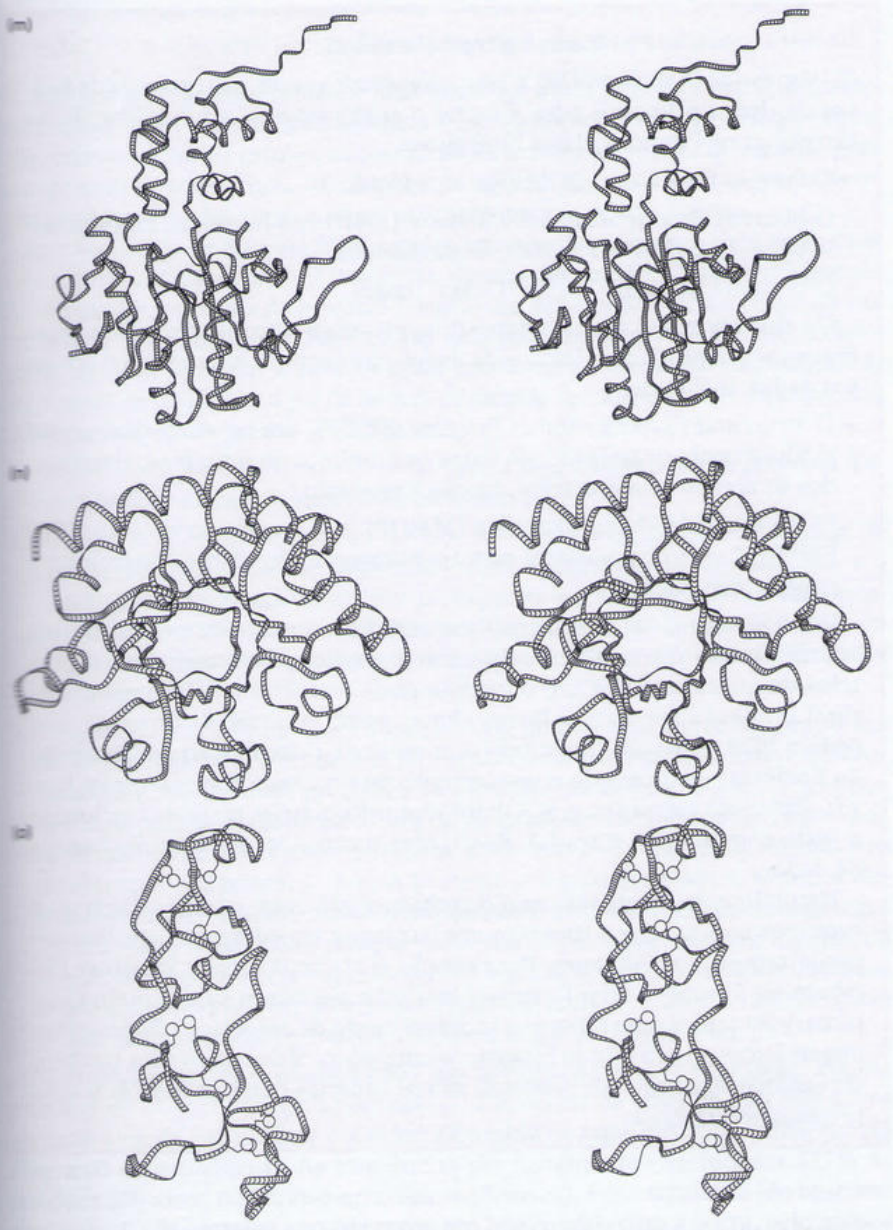


Figura 1.10 Um álbum de estruturas de proteínas. (a) Homeodomínio [código PDB: 1ENH]. (b) Segundo domínio homólogo calponina da utrofina [código PDB: 1BHD]. (c) Domínio de ligação ao DNA da proteína HIN recombinase (proteína em preto, DNA em vermelho) [código PDB: 1HCR]. (d) Citocromo *c* do embrião de arroz [código PDB: 1CCR]. (e) Módulo de adesão celular tipo III-10 da fibronectina [código PDB: 1FNA]. (f) Aglutinina manose específica (lectina) [código PDB: 1NPL]. (g) Domínio principal da proteína TBP (TATA-Box-binding protein – preto) de ligação ao DNA (vermelho) [código PDB: 1CDW]. (h) Barnase [código PDB: 1BRN]. (i) Lisil-tRNA sintetase [código PDB: 1BBW]. (j) Citalone desidratase [código PDB: 3STD]. (k) Domínio de ligação ao NAD da álcool desidrogenase. NAD em vermelho. [código PDB: 1EE2]. (l) Adenilato cinase [código PDB: 3ADK]. (m) Receptor quimiotático da metiltransferase [código PDB:1AF7]. (n) Tiamina fosfato sintase [código PDB: 2TPS]. (o) Polipeptídeo espasmolítico pancreático de suínos [código PDB: 2PSP].

Recursos na *web*: Estruturas macromoleculares

O Worldwide PDB (wwPDB) é uma colaboração entre três projetos de bancos de dados primários para integrar o arquivamento e a distribuição de estruturas macromoleculares biológicas:

- O Protein Data Bank (PDB) (Estados Unidos)
- O Macromolecular Structure Database (MSD) (no European Bioinformatics Institute (EBI) – Hinxton, Reino Unido)
- O Protein Data Bank /Japan (Osaka, Japão)

Os sítios wwPDB aceitam depósitos, processam novas entradas e mantêm os arquivos. Outros bancos de dados reorganizam e fornecem o acesso aos dados, incluindo:

- O Structural Classification of Proteins (SCOP), um banco de dados, meticulosamente organizado, de todos os domínios de proteínas, classificados de acordo com estrutura, função e evolução.
- O Molecular Modeling Database (MMDB), um projeto do sistema NCBI ENTREZ, que disponibiliza estruturas macromoleculares determinadas experimentalmente.

Estes e muitos outros sítios fornecem ferramentas de pesquisa para identificar estruturas de interesse. Por exemplo, para localizar uma proteína de interesse no SCOP, o usuário pode percorrer a hierarquia estrutural ou pesquisar com palavras-chave, como o nome da proteína, seu código PDB, a função (incluindo o número de classificação da Comissão de Enzimas) e o nome do enovelamento ou topologia (por exemplo, barril). Para cada estrutura, o SCOP fornece informações textuais (incluindo o texto completo da entrada), ilustrações e conexões para outros bancos de dados.

Naturalmente, existe uma grande sobreposição entre os sítios. Cada qual com seus pontos fortes, baseados, muitas vezes, no interesse científico dos pesquisadores contribuintes. Por exemplo, o Macromolecular Structure Database no European Bioinformatics Institute mantém o sítio Protein Quaternary Structure, que fornece o provável modo de associação de proteínas oligoméricas no seu estado biologicamente ativo. Sítios diferentes também têm aparência e utilidade distintas; são os usuários que irão descobrir suas próprias preferências.

Acompanhe o percurso das cadeias visualmente, reconhecendo hélices e folhas β. (As listras anguladas, preenchendo as estruturas, indicam a direção da cadeia.) Você consegue ver estruturas supersecundárias? Em quais das classes estruturais estas proteínas podem ser agrupadas? (Ver Exercícios 1.13 e 1.14 e Problema 1.2.) Muitos outros exemplos aparecem nos livros *Introduction to Protein Architecture: The Structural Biology of Proteins* e *Introduction to Protein Science: Architecture, Function and Genomics*.

Predição e engenharia da estrutura de proteínas

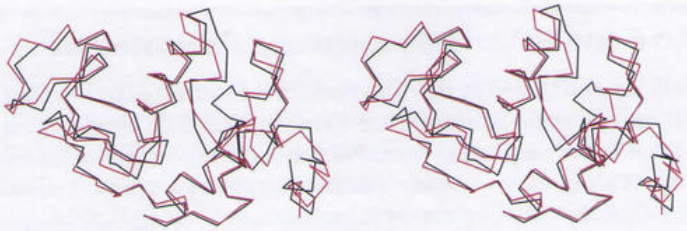
A seqüência de aminoácidos de uma proteína determina sua estrutura tridimensional. Em um meio com solvente e em condições de temperatura adequados, como o interior celular, as proteínas enovelam-se espontaneamente para atingir seus estados ativos. As chaperonas ajudam as proteínas a se enovelarem corretamente, porém catalisando o processo em vez de conduzi-lo.

Se a seqüência de aminoácidos contém informação suficiente para especificar a estrutura tridimensional de proteínas, deveria ser possível o desenvolvimento de um algoritmo capaz de predizer a estrutura de uma proteína a partir de sua seqüência de aminoácidos. Isso tem se mostrado bastante difícil. Como conseqüência, além de tentar a solução deste problema fundamental da predição de estruturas de proteínas a partir de sua seqüência de aminoácidos, os cientistas definiram objetivos menos ambiciosos:

1. **Predição da estrutura secundária** Quais segmentos de uma proteína formam hélices e quais formam fitas ou folhas β?
2. **Reconhecimento de padrão de enovelamento** Dada uma biblioteca de estruturas conhecidas de proteínas e suas seqüências de aminoácidos, e a seqüência de aminoácidos de uma proteína de estrutura desconhecida, podemos encontrar a estrutura na biblioteca que tenha a maior probabilidade de apresentar um padrão de enovelamento similar ao da proteína de estrutura desconhecida?
3. **Modelagem por homologia** Suponha que uma proteína-alvo, de seqüência de aminoácidos conhecida mas de estrutura indeterminada, esteja relacionada com uma ou mais proteínas de estrutura conhecida. Então, esperamos que uma grande parte da estrutura da proteína-alvo se pareça com aquela da proteína conhecida, e esta, portanto, pode servir de base para um modelo da estrutura da proteína-alvo. A precisão e a qualidade dos resultados dependem fundamentalmente da similaridade entre as seqüências. Como regra prática, se as seqüências de duas proteínas homólogas possuem identidade entre resíduos igual ou superior a 50% em um alinhamento ótimo, as suas estruturas provavelmente serão similares em mais de 90% do modelo. (Esta é uma estimativa conservadora, como mostra a ilustração a seguir.)

Aqui podemos ver o alinhamento das seqüências e a sobreposição das estruturas de duas proteínas relacionadas, a lisozima da clara de ovo (em preto no diagrama de estruturas) e a α-lactalbumina de babuínos (em vermelho no diagrama de estruturas). As seqüências são fortemente relacionadas (37% de resíduos idênticos no alinhamento das seqüências), e as estruturas são bastante similares. Cada proteína serviria como um bom molde para a outra, pelo menos no que diz respeito à orientação da cadeia principal.

```
Lisozima de galinha          KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGS
α-lactalbumina de babuíno    KQFTKCELSQNLY--DIDGYGRIALPELICTMFHTSGYDTQAIVEND-ES

Lisozima de galinha          TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS
α-lactalbumina de babuíno    TEYGLFQISNALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKILD

Lisozima de galinha          DGN-GMNAWVAWRNRCKGTDVQA-WIRGCRL-
α-lactalbumina de babuíno    I--KGIDYWIAHKALC-TEKL-EQWL--CE-K
```



Critical Assessment of Structure Prediction (CASP)

A avaliação das técnicas de predição de estruturas de proteínas é realizada por meio de testes às cegas. Com este objetivo, J. Moult iniciou os programas CASP (*Critical Assessment of Structure Prediction*) bianuais. Cristalógrafos e espectroscopistas de NMR a caminho da determinação da estrutura de uma proteína são convidados a (1) publicarem a seqüência de aminoácidos das proteínas vários meses antes da data esperada para a conclusão de seus experimentos e (2) se comprometerem em manter seus dados em sigilo até uma data marcada. Grupos que trabalham com predição de estruturas submetem seus modelos, os quais são mantidos até a liberação da estrutura experimental. Então, os dados de predição e os dados experimentais são comparados – para o deleite de poucos e o aborrecimento da maioria.

Os resultados das avaliações do CASP registram o progresso na eficácia das predições, o que ocorreu em parte por causa do crescimento dos bancos de dados, mas, também, como conseqüência dos aprimoramentos nos métodos de predição. Discutiremos a predição de estrutura de proteínas no Capítulo 5.

Engenharia de proteínas

Os biólogos moleculares costumavam ser como os astrônomos – podíamos observar nossos objetos de estudo, mas não modificá-los. Isso não é mais a realidade. É possível manipular aminoácidos e ácidos nucléicos à vontade em laboratórios. Podemos testá-los por meio de mutações exaustivas para ver os efeitos na função. Podemos dar novas funções a velhas proteínas, como no desenvolvimento de anticorpos catalíticos. Podemos, até mesmo, criar novas proteínas.

Muitas das regras sobre a estrutura de proteínas foram derivadas da observação de proteínas de ocorrência natural. Essas regras não se aplicam, *necessariamente*, às proteínas engenhadas. Proteínas naturais têm características determinadas pelos princípios gerais da físico-química e pelo mecanismo de evolução das proteínas. As proteínas engenhadas devem obedecer aos princípios físico-químicos, mas não às restrições da evolução. Com as proteínas engenhadas, podemos explorar novos territórios.

Proteômica

O proteoma, em analogia ao genoma, é o conjunto de proteínas de um organismo. A proteômica combina identificação, distribuição, interações, dinâmica e padrões de expressão das proteínas de sistemas vivos. R. Simpson criou a analogia: se o genoma é uma lista dos instrumentos em uma orquestra, as proteínas são a orquestra executando uma sinfonia. É um assunto que envolve um grande volume de informações e que depende de técnicas rápidas de coleta de dados em larga escala. Entre essas técnicas encontram-se a análise com microarranjos de DNA e a espectrometria de massa.

Microarranjos de DNA

Ver Quadro Aplicações de microarranjos de DNA.

Microarranjos de DNA, ou *chips* de DNA, são instrumentos para testar, *simultaneamente*, a presença de muitas seqüências de DNA. Microarranjos de DNA podem ser utilizados para (1) determinar padrões de expressão de diferentes proteínas pela detecção de mRNAs; ou (2) genotipar, pela detecção de diferentes seqüências de genes variantes, incluindo mas não se limitando aos polimorfismos de um único nucleotídeo (SNPs). É possível medir uma simples presença ou ausência de SNPs, ou quantificar a abundância relativa.

Do ponto de vista da bioinformática, os microarranjos de DNA são, de fato, outra fonte prolífica de geração de dados. Eles demandam planejamento eficaz de arquivos de dados e de sistemas de processamento de informações. Uma vantagem é que os dados são todos tão novos que este campo de pesquisa não está sobrecarregado com estruturas de dados e formatos baseados em gerações mais antigas de *hardware* e programas de computadores.

Aplicações de microarranjos de DNA

- **Identificação da individualidade genética em tecidos ou organismos** Em humanos e animais, microarranjos de DNA permitem a correlação entre genótipo e suscetibilidade a doenças. Em bactérias, permitem a identificação dos mecanismos de desenvolvimento de resistência a fármacos por patógenos.
- **Investigação de estados e processos celulares** Padrões de expressão que mudam com o estado da célula ou condições de crescimento podem fornecer pistas sobre os mecanismos de esporulação, ou sobre a mudança entre os estados aeróbios e anaeróbios do metabolismo.
- **Diagnóstico de doenças** Testes para a presença de mutações podem confirmar o diagnóstico de uma doença genética suspeita. A detecção de portadores pode auxiliar no aconselhamento familiar.
- **Sinais de alerta genéticos** Algumas doenças não são determinadas de forma total e irrevogável pelo genótipo, mas pela probabilidade de o seu desenvolvimento estar relacionado com genes ou seus padrões de expressão. Uma pessoa consciente do maior risco de desenvolver uma doença pode, em alguns casos, melhorar seu prognóstico com ajustes em seu estilo de vida ou, em outras doenças, com cirurgias profiláticas.
- **Seleção de fármacos** Detecção de fatores genéticos que determinam respostas a fármacos, que em alguns pacientes tornam o tratamento ineficiente, e em fármacos causam reações adversas incomuns graves.
- **Diagnóstico especializado de doenças** Tipos diferentes de leucemia podem ser identificados por padrões diferentes de expressão gênica. O conhecimento do tipo exato de doença é importante para o prognóstico e para a escolha do tratamento adequado.
- **Seleção de alvos para o planejamento de fármacos** Proteínas que apresentam transcrição exacerbada, em condições particulares de uma doença, podem ser boas candidatas para tentativas de intervenção farmacológica.

→

Aplicações de microarranjos de DNA (*continuação*)

- **Resistência de patógenos** A comparação de genótipos ou de padrões de expressão, entre cepas de bactérias suscetíveis e resistentes a um antibiótico, indica as proteínas envolvidas nos mecanismos de resistência.
- **Acompanhamento das variações temporais no padrão de expressão de proteínas** Este acompanhamento permite medir (1) o tempo de resposta a infecções por patógenos, (2) respostas a alterações ambientais e (3) mudanças que ocorrem durante o ciclo celular.

Espectrometria de massa

A espectrometria de massa é uma técnica física que caracteriza moléculas pela mensuração das massas de seus íons. As aplicações na área da proteômica incluem:

- A identificação rápida de componentes de uma mistura complexa de proteínas.
- O seqüenciamento de proteínas e ácidos nucléicos.
- A análise de modificações pós-traducionais, ou de substituições relativas a uma seqüência esperada.
- O monitoramento da troca hidrogênio-deutério para revelar a exposição ao solvente de diferentes sítios da proteína. Isso fornece informações sobre a conformação estática e dinâmica – incluindo enovelamento e interações.

Biologia de sistemas

O lema da biologia de sistemas é integração. Biólogos moleculares passaram um século estudando as células separadamente – purificando proteínas individualmente e avaliando suas propriedades isoladamente. Nosso trabalho agora é unir todas estas informações novamente.

A integração possui dois aspectos. Um é o estudo de padrões *em uma célula ou em um organismo*: padrões de interação proteína-proteína e proteína-ácidos nucléicos, padrões de vias metabólicas e cascatas de controles, e padrões de expressão de proteínas. Os padrões possuem tanto aspectos estáticos quanto dinâmicos. A identificação de pares de proteínas que se ligam uma à outra e o agrupamento das interações entre esses pares de proteínas em uma via produzem um padrão estático. O fluxo de metabólitos através de uma rede de enzimas ou o fluxo de informação ao longo de uma cascata de controle são padrões dinâmicos.

O outro aspecto da integração é a comparação de ocorrência, atividades e interações de genes e proteínas *através de espécies diferentes*. A razão pela qual a metodologia comparativa é tão poderosa na biologia é que estamos tentando entender sistemas que surgiram por meio de processos evolucionários. Espécies diferentes podem esclarecer fatos umas das outras. Para compreender o que significa ser humano, devemos apreciar tanto o que temos em comum com outras espécies quanto o que nos diferencia delas.

Técnicas rápidas de produção de dados em larga escala em genômica e proteômica fornecem informações sobre seqüências, padrões de expressão e de interações. A partir da seqüência de genomas, podemos inferir as seqüências de aminoácidos de todas as proteínas de um organismo. A proteômica nos diz como os padrões de expressão dessas proteínas variam em um organismo, como eles se alteram durante o desenvolvimento ou em resposta a mudanças nas condições

ambientais, e como as proteínas cooperam entre si. A biologia de sistemas une essas informações como as peças de um quebra-cabeça que se estende no tempo e no espaço. Para compreender este agente complexo e delicado que é uma célula viva, precisamos encaixar todas as peças em seus devidos lugares.

Implicações clínicas

Existe um consenso de que o seqüenciamento do genoma humano e de outras espécies vai melhorar a saúde da humanidade. Mesmo desconsiderando as pretensões mais exageradas – a publicidade sensacionalista dura uma eternidade – as categorias de aplicações incluem:

1. **Diagnóstico de doenças e de riscos de doenças** O seqüenciamento do DNA pode detectar a ausência de um gene particular, ou de uma mutação. A identificação de seqüências gênicas específicas associadas a doenças permitirá diagnósticos rápidos e confiáveis de estados de saúde precários (a) quando o paciente apresentar sintomas, (b) antes que os sintomas apareçam, como em testes para condições hereditárias de manifestação tardia, como a doença de Huntington (ver Quadro, página 76), (c) para o diagnóstico *in utero* de potenciais anormalidades como a fibrose cística, e (d) para o aconselhamento genético de casais que pretendem ter filhos.

   Em muitos casos, os nossos genes não nos condenam ao desenvolvimento de uma dada doença de maneira irrevogável, mas aumentam a probabilidade de que isso ocorra. Um exemplo de fator de risco detectável em nível genético envolve a α_1-antitripsina, uma proteína que normalmente atua na inibição da elastase nos alvéolos pulmonares. Indivíduos homozigotos para o mutante Z da α_1-antitripsina (342Glu→Lis) expressam apenas uma proteína disfuncional. Esses indivíduos têm maior risco de desenvolver enfisema, por causa dos danos causados aos pulmões pela elastase endógena que, normalmente, seria regulada pela atividade inibitória, e também de doenças hepáticas, causadas pelo acúmulo da forma polimérica da α_1-antitripsina nos hepatócitos onde são sintetizadas. O hábito de fumar favorece, com certeza, o desenvolvimento de enfisema. No caso de indivíduos homozigotos para o mutante Z, a doença se desenvolve como uma *combinação* de fatores genéticos e ambientais.

   "A genética carrega a arma e o ambiente puxa o gatilho" – J. Stern

   Geralmente a relação entre o genótipo e fatores de risco de doenças é muito mais difícil de ser determinada. Algumas doenças, como a asma, dependem das interações de muitos genes, assim como de fatores ambientais. Em outros casos, o gene pode estar presente e íntegro, mas mutações em outros locais do genoma podem alterar seus níveis de expressão ou sua distribuição nos tecidos. Essas anomalias devem ser detectadas pela medição da atividade protéica. A análise de padrões de expressão de proteínas também é uma forma importante de medir a resposta ao tratamento.

2. **A genética de respostas à terapia – tratamento personalizado** Uma vez que as pessoas diferem em suas capacidades de metabolizar fármacos, pacientes diferentes com a mesma doença podem necessitar de dosagens diferentes. A análise de seqüências permite selecionar fármacos e dosagens ótimas para cada paciente, um campo de pesquisa em rápido desenvolvimento denominado **farmacogenômica**. Os médicos podem, assim, evitar a experimentação de diferentes terapias, um procedimento que é perigoso em termos de efeitos colaterais – muitas vezes fatal – e, em todo caso, oneroso. O tratamento de pacientes por reações adversas a fármacos prescritos custa bilhões de dólares aos sistemas de saúde.

Por exemplo, a 6-mercaptopurina, um fármaco extremamente tóxico, é utilizado no tratamento de leucemia infantil. Uma pequena fração dos pacientes costumava morrer em função desse tratamento, pois não possuía a enzima tiopurina metiltransferase, necessária para metabolizar o fármaco. O teste de pacientes para a presença ou ausência dessa enzima identifica aqueles sob risco.

De maneira inversa, pode também se tornar possível o uso de fármacos que são seguros e eficazes em uma minoria de pacientes, mas que foram rejeitados antes ou durante os testes clínicos por não serem eficazes, ou por apresentarem efeitos colaterais na maioria dos pacientes.

A doença de Huntington

A doença de Huntington é um distúrbio neurodegenerativo hereditário que afeta aproximadamente 30.000 pessoas nos Estados Unidos. Seus sintomas são bastante graves, incluindo movimentos involuntários semelhantes a passos de dança (coréicos), transtornos mentais, mudanças de personalidade e diminuição da capacidade intelectual. A morte ocorre cerca de 10 a 15 anos após o início dos sintomas. O gene surgiu na Nova Inglaterra durante o período colonial, no século XVII. Alguns casos de bruxaria da época podem estar relacionados à doença. O gene não foi eliminado da população, pois a idade em que se manifesta – entre 30 e 50 anos – ocorre após o período reprodutivo típico.

Antigamente, membros de famílias afetadas não tinham alternativa a não ser encarar a incerteza e o medo, durante a juventude e o início da vida adulta, de não saber se haviam herdado a doença. A descoberta do gene da doença de Huntington em 1993 tornou possível identificar indivíduos afetados. O gene contém repetições expandidas do trinucleotídeo CAG, correspondentes a blocos de poliglutamina na proteína correspondente, a *huntingtina*. (A doença de Huntington é uma de uma família de doenças neurodegenerativas que resultam da repetição de trinucleotídeos.) Quanto maiores os blocos de CAG, mais cedo os sintomas se manifestam e em maior intensidade. O gene normal contém entre 11 e 28 repetições CAG. Pessoas que apresentam entre 29 e 34 repetições têm baixa probabilidade de desenvolver a doença, e aquelas que apresentam entre 35 e 41 repetições podem desenvolver sintomas relativamente moderados. Entretanto, pessoas com mais de 41 repetições quase certamente sofrerão da doença de Huntington em sua forma mais grave.

A herança é marcada pelo fenômeno chamado de antecipação: as repetições aumentam em gerações sucessivas, aumentando progressivamente a gravidade da doença e reduzindo a idade de sua manifestação. Por alguma razão, este efeito é maior em genes paternos do que em genes maternos. Dessa forma, mesmo pessoas que ficam na região limítrofe, que podem carregar um gene que contenha entre 29 e 41 repetições, deveriam ser aconselhadas sobre os riscos para os seus filhos.

3. **Identificação de alvos para fármacos** Um **alvo** é uma proteína cuja função pode ser modificada de maneira seletiva pela interação com um fármaco para afetar os sintomas ou as causas de uma doença. A identificação de um alvo define o foco dos passos subseqüentes no processo de planejamento de fármacos. Dentre os fármacos em uso atualmente, o alvo de cerca da metade são receptores, perto de um quarto são enzimas e quase um quarto são hormônios. Aproximadamente 7% atuam em alvos desconhecidos.

O crescimento da resistência de bactérias ao tratamento com antibióticos está gerando uma crise no controle de doenças. Há uma possibilidade bastante real de que nossos descendentes venham a olhar para a segunda metade do século XX como um curto período de tempo durante o qual as infecções bacterianas podiam ser controladas e, antes e depois do qual, elas não podiam.

A urgência em encontrar novos fármacos é mitigada pela crescente disponibilidade de dados que podem servir de base para o seu desenvolvimento. A genômica pode sugerir alvos. A genômica diferencial e a comparação de padrões de expressão de proteínas entre cepas de bactérias patogênicas resistentes e suscetíveis a fármacos podem apontar as proteínas responsáveis pela resistência a fármacos. O estudo de alteração na expressão gênica entre células normais e células de tumores pode identificar proteínas expressas diferencialmente como alvos potenciais para fármacos anticâncer.

4. Terapia gênica Se um gene está ausente ou defeituoso, gostaríamos de poder restituí-lo, ou ao menos suprir o seu produto. Se um gene está superativado, gostaríamos de poder desativá-lo.

O suprimento direto de proteínas é possível em muitas doenças, das quais a reposição de insulina no diabetes e a de fator VIII para uma forma comum de hemofilia talvez sejam os exemplos mais conhecidos.

A transferência genética em animais foi bem-sucedida para a produção de proteínas humanas no leite de ovelhas e de vacas. Em pacientes humanos, a terapia de reposição gênica para casos de fibrose cística, utilizando adenovírus, mostrou resultados encorajadores.

Um método para o bloqueio de genes é chamado de *terapia anti-senso*. A idéia é introduzir uma seqüência curta de DNA ou de RNA que se liga de forma seqüência-específica em uma região do gene. A ligação ao DNA endógeno pode interferir com a sua transcrição; a ligação ao mRNA pode interferir com a sua tradução. A terapia anti-senso mostrou alguma eficácia contra o citomegalovírus e a doença de Crohn.

A terapia anti-senso é bastante atrativa, pois, indo diretamente ao bloqueio da seqüência-alvo, diminui muitas etapas do processo de planejamento de fármacos.

O futuro

O novo século verá uma revolução no desenvolvimento e no oferecimento de sistemas de amparo à saúde. As barreiras entre a pesquisa teórica e a prática clínica estão diminuindo. É possível que um leitor deste livro descubra a cura para uma doença que poderia matá-lo. Inclusive, é bastante provável que a frase espirituosa de Szent-Gyorgi, "O câncer ajuda mais as pessoas do que mata" se torne verdadeira. Espera-se que isso aconteça porque as instituições de pesquisa tiveram sucesso no desenvolvimento de medidas terapêuticas ou preventivas contra tumores em vez de simplesmente imitar o seu crescimento descontrolado.

Leituras recomendadas

Um vislumbre do futuro?

Blumberg, B. S. (1996), Medical research for the next millenium, *The Cambridge Review*, **117**, 3-8. [Uma predição fascinante do que está por vir, e do que já está acontecendo.]

O cenário intelectual

Mayr, E., *What Makes Biology Unique? Considerations on the Autonomy of a Scientific Discipline*. (Cambridge: Cambridge University Press, 2004). [Perspectivas da biologia, por um cientista que se autodescreveu como "um biólogo de unhas sujas", com uma clareza de pensamento sem igual.]

Sobre o contexto geral da biologia

Doolittle, W. F. (2000), Uprooting the tree of life, *Sci. Am.*, **282(2)**, 90-95. [Implicações da análise de seqüências para a nossa compreensão das relações entre os organismos vivos.]

Determinação de seqüências genômicas

Green, E. D. (2001), Strategies for systematic sequencing of complex organisms, *Nature Reviews (Genetics)*, **2**, 573-583. [Uma discussão clara de possíveis métodos para projetos de seqüenciamento em larga escala. Inclui listas de, e conexões para, projetos em andamento de seqüenciamento de organismos multicelulares.]

Sulston, J. & Ferry, G., *The Common Thread: a story of science, politics, ethics and the human genome* (New York: Bantam, 2002). [Um relato de primeira mão e pessoal sobre uma das maiores operações científicas internacionais já realizadas, o seqüenciamento do genoma humano.]

Discussões sobre bancos de dados e recuperação de informações

Altschul, S. F., Madden, T. L., SchŠffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402.

Frishman, D., Heumann, K., Lesk, A. & Mewes, H. -W. (1998), Comprehensive, comprehensible, distributed and intelligent databases: Current status, *Bioinformatics*, **14**, 551-561.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A. & Wagner, L. (2003), Database resources of the National Center for Biotechnology, *Nucleic Acids Res.*, **31**, 28-33.

Lesk, A. M. and 25 co-authors, (2000), Quality control in databanks for molecular biology, *BioEssays*, **22**, 1024-1034.

Stein, L. (2001), Genome annotation: from sequence to biology, *Nature Reviews (Genetics)*, **2**, 493-503.

Lesk, A. M., Editor, *Database Annotation in Molecular Biology: Principles and Practice* (Chichester: J. Wiley and Sons, 2004). [Um conjunto de artigos descrevendo o quê os bancos de dados requerem no que diz respeito à anotação, e quais tipos de habilidades são necessárias para os especialistas em anotação.]

Proteínas

Branden, C. I. & Tooze, J., *Introduction to Protein Structure* (2nd edn. New York: Garland, 1999). [Um texto introdutório de excelente qualidade.]

Lesk, A. M., *Introduction to Protein Architecture: The Structural Biology of Proteins* (Editora original deste livro, 2000).

Lesk, A. M., *Introduction to Protein Science: Architecture, Function and Genomics* (Editora original deste livro, 2004).

[Volumes que acompanham o livro *Introdução à Bioinformática*, com foco em estrutura, função e evolução de proteínas.]

A transição para a publicação eletrônica

Lesk. M., *Understanding Digital Libraries* (2nd edn. San Francisco: Morgan Kaufmann, 2004). [Introdução à transição de bibliotecas tradicionais para a provisão de informações pelo computador.]

Berners-Lee, T. (with Mark Fischetti), *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web* (New York: HarperBusiness, 2000).

Berners-Lee, T. & Hendler, J. (2001), Publishing on the semantic web, *Nature*, **410**, 1023-1024. [Um artigo escrito pelo inventor da *web*.]

Butler, D. & Campbell, P. (2001), Future e-access to the primary literature, *Nature*, **410**, 613. [Um artigo que descreve os desenvolvimentos na publicação eletrônica de periódicos científicos.]

Malakoff, D. (2003), Scientific publishing Opening the books on open access, *Science*, **302**, 550-554. [Uma descrição dos periódicos publicados pela Public Library of Science (PLoS).]

Spedding V. (2003), Great data, but will it last? *Research Information* (Spring 2003) 16-20. [Artigo abordando os problemas de preservação da informação digital. Este periódico possui muitos artigos de interesse para cientistas cuja pesquisa depende da qualidade dos dados e sua acessibilidade por meio de computadores.]

Aspectos legais

Human Genome Project Information Website: Genetics and Patenting
`http://www.ornl.gov/hgmis/elsi/patents.html`

Bobrow, M. & Thomas, S. (2002), Patenting DNA, *Curr. Opin. Mol. Ther.*, **4**, 542–547.

F. S. Kieff, ed. (2003), Perspectives on Properties of the Human Genome Project, *Adv. Genetics*, vol. **50**. [Uma coleção de artigos discutindo aspectos legais de genômica e bioinformática.]

Exercícios, problemas e *weblemas*

Exercícios

1.1 (a) O Sloan Digital Sky Survey é um mapeamento do céu do hemisfério norte por um período de cinco anos. O total de dados brutos chegará a cerca de 15 terabytes (1 byte = 1 caractere, 1TB = 10^{12} bytes). A quantos projetos genoma este volume de dados equivale? (b) O Earth Observing System/Data Information System (EOS/DIS) – uma série de observações globais de longo prazo da Terra – exige aproximadamente 15 petabytes de dados armazenados (1 petabyte = 10^{15} bytes). A quantos projetos genoma este volume de dados equivale? (c) Compare o volume para armazenamento de dados necessários para o EOS/DIS com o volume de dados necessários para armazenar as seqüências completas de DNA de todos os habitantes dos Estados Unidos. (Ignore os meios de compressão de dados utilizados para o armazenamento de arquivos. Suponha que a seqüência de DNA de cada pessoa exija 1 byte/nucleotídeo.)

1.2 (a) Quantos disquetes seriam necessários para armazenar o genoma humano completo? (b) Quantos CDs seriam necessários para armazenar o genoma humano completo? (c) Quantos DVDs seriam necessários para armazenar o genoma humano completo? (Em todos os casos, suponha que a seqüência seja armazenada a um byte por caractere, sem compressão.)

1.3 Suponha que você vai preparar o quadro sobre a doença de Huntington (página 76) para um sítio na *web*. Para quais palavras ou frases você forneceria conexões?

1.4 O final do gene β-hemoglobina humano possui a seqüência de nucleotídeos:

```
... ctg gcc cac aag tat cac taa
```

(a) Qual é a tradução desta seqüência em uma seqüência de aminoácidos? (b) Escreva a seqüência de nucleotídeos com uma única alteração de base que produza uma mutação silenciosa nesta região. (Uma mutação silenciosa é aquela que não altera a seqüência de aminoácidos.) (c) Escreva a seqüência de nucleotídeos, e a tradução em seqüência de aminoácidos, com a alteração de uma única base, produzindo uma mutação não-sinônima na região. (d) Escreva a seqüência de nucleotídeos, e a tradução em seqüência de aminoácidos, com a alteração de uma única base, produzindo uma mutação nesta região que levaria ao truncamento precoce da proteína. (e) Escreva a seqüência de nucleotídeos com a alteração de uma única base, produzindo uma mutação nesta região que levaria à terminação imprópria da cadeia, originando uma extensão da proteína.

1.5 Em uma fotocópia do Quadro Alinhamento de seqüência par a par completo das proteínas PAX-6 humana e *eyeless* de *Drosophila melanogaster*, indique com uma caneta marca-texto as regiões alinhadas pelo PSI-BLAST.

1.6 Em uma cópia do alinhamento da proteína PAX-6 humana e *eyeless* de *Drosophila melanogaster* (página 57), marque as regiões alinhadas à proteína *even-skipped* pelo PSI-BLAST.

1.7 (a) Que valores de corte de *E* você usaria em uma pesquisa com o PSI-BLAST se tudo o que você deseja saber é se a sua seqüência já está no banco de dados? (b) Que valores de corte de *E* você usaria em uma pesquisa com o PSI-BLAST se você deseja localizar homólogos distantes de sua seqüência?

1.8 No desenho de uma seqüência anti-senso, estime o comprimento mínimo necessário para evitar uma complementaridade exata com regiões aleatórias do genoma humano.

1.9 Sugere-se que todos os seres humanos vivos sejam descendentes de um ancestral comum chamado Eva, que teria vivido entre aproximadamente 140.000 e 200.000 anos atrás. (a) Assumindo que haja seis gerações por século, quantas gerações existem entre Eva e o presente? (b) Se uma célula bacteriana se divide a cada 20 minutos, quanto tempo levaria para a bactéria passar por este mesmo número de gerações?

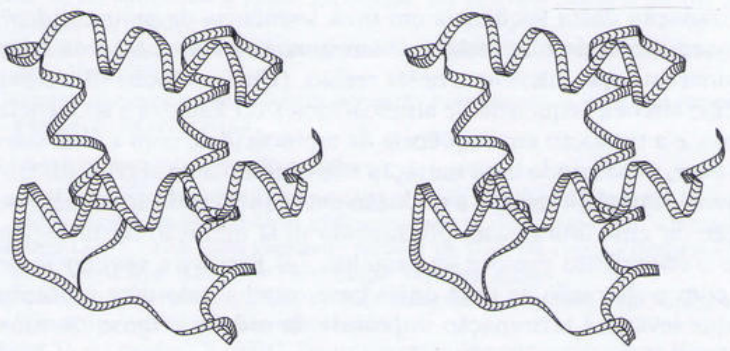
1.10 Indique um aminoácido que tenha características físico-químicas semelhantes (a) à leucina, (b) ao ácido aspártico, (c) à treonina. Esperamos que tais substituições, na maioria dos casos, tenham efeitos relativamente pequenos na estrutura e função de uma proteína. Indique um aminoácido que tenha características físico-químicas bastante distintas (a) da leucina, (b) do ácido aspártico, (c) da treonina. Essas substituições deveriam ter um efeito drástico na estrutura e função de uma proteína, especialmente se ocorrerem no seu interior.

1.11 Na Figura 1.7 (a), a direção da cadeia no sentido N-terminal para C-terminal aponta para cima ou para baixo na página? Na Figura 1.7 (b), a direção da cadeia no sentido N-terminal para C-terminal aponta para cima ou para baixo na página?

1.12 Analisando a Figura 1.9, quantas vezes a cadeia passa entre os domínios da proteína ribossomal L1 de *M. jannaschii*?

1.13 Em uma fotocópia da Figura 1.10 (k e l), indique com uma caneta marca-texto as hélices (em vermelho) e as fitas de folha (em azul). Em uma fotocópia da Figura 1.10 (g e m), divida a proteína em domínios.

1.14 Qual das estruturas mostradas na Figura 1.10 contém o seguinte domínio?



1.15 Em uma fotocópia da superposição das estruturas da lisozima de galinha e da α-lactalbumina de babuíno, indique com uma caneta marca-texto duas regiões nas quais as conformações das cadeias principais são diferentes.

1.16 No programa PERL da página 40, estime a fração do texto do programa que contém comentários (conte linhas inteiras e pela metade).

1.17 Modifique o programa PERL que extrai o nome das espécies do arquivo de saída do PSI-BLAST para que ele também aceite nomes no formato `[D. melanogaster]`.

1.18 Modifique o programa PERL que extrai o nome das espécies do arquivo de saída do PSI-BLAST para que ele conte o número de seqüências de cada espécie ocorrendo na lista.

1.19 Qual é a seqüência de nucleotídeos da molécula mostrada na Gravura I?

Problemas

1.1 A tabela seguinte contém um alinhamento múltiplo de seqüências parciais de uma família de proteínas chamadas domínios ETS. Cada linha corresponde à seqüência de aminoácidos de uma proteína, e cada letra corresponde a um aminoácido. As colunas indicam qual aminoácido está presente naquela posição em cada uma das proteínas na família. Dessa forma, padrões de preferências se tornam visíveis.

```
TYLWEFLLKLLQDR.EYCPRFIKWTNREKGVFKLV..DSKAVSRLWGMHKN.KPD
WQLWQFLLEILTD..CEHTDVIEWVG.TEGEFKLT..DPDRVARLWGEKKN.KPA
IQLWQFLLELLTD..KDARDCISWVG.DEGEFKLN..QPELVAQKWGQRKN.KPT
IQLWQFLLELLSD..SSNSSCITWEG.TNGEFKMT..DPDEVARRWGERKS.KPN
IQLWQFLLELLTD..KSCQSFISWTG.DGWEFKLS..DPDEVARRWGKRKN.KPK
IQLWQFLLELLQD..GARSSCIRWTG.NSREFQLC..DPKEVARLWGERKR.KPG
IQLWHFILELLQK..EEFRHVIAWQQGEYGEFVIK..DPDEVARLWGRRKC.KPQ
VTLWQFLLQLLRE..QGNGHIISWTSRDGGEFKLV..DAEEVARLWGLRKN.KTN
ITLWQFLLHLLLD..QKHEHLICWTS.NDGEFKLL..KAEEVAKLWGLRKN.KTN
LQLWQFLVALLDD..PTNAHFIAWTG.RGMEFKLI..EPEEVARLWGIQKN.RPA
IRLWQFLKELLASP.QVNGTAIRWIDRSKGIFKIE..DSVRVAKLWGRRKN.RPA
RLLWDFLQQLLNDRNQKYSDLIAWKCRDTGVFKIV..DPAGLAKLWGIQKN.HLS
RLLWDYVYQLLSD..SRYENFIRWEDKESKIFRIV..DPNGLARLWGNHKN.RTN
IRLYQFLLDLLRS..GDMKDSIWWVDKDKGTFQFSSKHKEALAHRWGIQKGNRKK
LRLYQFLLGLLTR..GDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKR
```

Em uma fotocópia desta página:

(a) Utilizando canetas marca-texto coloridas, destaque, em cada seqüência, com diferentes cores, os resíduos de diferentes classes:

resíduos pequenos:	G A S T
resíduos apolares de tamanho médio:	C P V I L
resíduos apolares grandes:	F Y M W
resíduos polares:	H N Q
resíduos carregados positivamente:	K R
resíduos carregados negativamente:	D E

(b) Para cada posição que contenha o mesmo aminoácido em todas as seqüências, escreva a letra que simboliza este aminoácido, em letra maiúscula, abaixo da coluna correspondente. Para cada posição que contenha o mesmo aminoácido em todas as proteínas, exceto uma, escreva a letra que simboliza este aminoácido, em letra minúscula, abaixo da coluna correspondente.

(c) O que o padrão de periodicidade dos resíduos conservados sugere?

(d) Que estruturas secundárias estes padrões sugerem em certas regiões?

(e) Que distribuição de conservação de resíduos carregados você observa? Proponha ("chute") o tipo de molécula com a qual estes domínios interagem.

1.2 Classifique as estruturas da Figura 1.10 de acordo com as seguintes categorias: hélices α, folhas β, α+β, α/β linear, barril α/β, pouca ou nenhuma estrutura secundária.

1.3 Generalize o programa PERL da página 38 para imprimir a tradução da seqüência de DNA em todas as seis fases de leitura possíveis.

1.4 Escreva um programa PERL para ler um alinhamento do CLUSTAL-W, tal como o alinhamento do citocromo *b* mitocondrial de elefantes e mamute, e contar o número de malpareamentos entre todos os pares de proteínas.

1.5 Para qual dos seguintes conjuntos de cadeias de fragmentos o programa PERL da página 40 funciona corretamente?

(a) Ele recuperaria corretamente o texto:

Kate, when France is mine and I am
yours, then yours is France and you are mine.

a partir de:

Kate, when France
France is mine
is mine and
and I am\nyours
yours then
then yours is France
France and you are mine\n

(b) Ele recuperaria corretamente o texto:

One woman is fair, yet I am well; another is wise, yet I am well; another virtuous, yet I am well; but till all graces be in one woman, one woman shall not come in my grace.

a partir de:

One woman is
woman is fair,
is fair, yet I am
yet I am well;
I am well; another
another is wise, yet I am well;
yet I am well; another virtuous,
another virtuous, yet I am well;
well; but till all
all graces be
be in one woman,
one woman, one
one woman shall
shall not come in my grace.

(c) Ele recuperaria corretamente o texto:

That he is mad, 'tis true: 'tis true 'tis pity;
And pity 'tis 'tis true.

a partir de:

That he is
is mad, 'tis
'tis true
true: 'tis true 'tis
true 'tis
'tis pity;\n
pity;\n And pity
pity 'tis
'tis 'tis
'tis true.\n

Em (c), o programa funcionaria se toda a pontuação fosse removida das cadeias?

1.6 Generalize o programa PERL da página 40 para que ele agrupe corretamente todos os fragmentos de textos do problema anterior. (Atenção: este não é um problema fácil.)

1.7 Escreva um programa PERL que encontre alinhamento de motivos como ilustrado no Quadro da página 46. (a) Exija alinhamento perfeito ou exato. (b) Permita um malpareamento, não necessariamente nas primeiras posições, como no exemplo, mas não permita inserções ou deleções.

1.8 A programação em PERL pode ser bastante concisa. Esta é uma versão alternativa do programa para montar fragmentos com sobreposição (ver página 40):

```
#!/usr/bin/perl

$/ = "";
@fragments = split("\n",<DATA>);

foreach (@fragments) { $firstfragment{$_} = $_; }

foreach $i (@fragments) {
    foreach $j (@fragments) { unless ($i eq $j) {
        ($combine = $i . "XXX" . $j) =~ /([\S ]{2,})XXX\1/;
        (length($1) <= length($successor{$i})) || { $successor{$i} = $j };
    }                          }
    undef $firstfragment{$successor{$i}};
}

$test = $outstring = join "", values(%firstfragment);
while ($test = $successor{$test}) { ($outstring .= "XXX".$test) =~ s/([\S ]+)XXX\1/\1/;}

$outstring =~ s/\\n/\n/g; print "$outstring\n";

__END__
the men and women merely players;\n
one man in his time

All the world's
their entrances,\nand one man
stage,\nAnd all the men and women
They have their exits and their entrances,\n
world's a stage,\nAnd all
their entrances,\nand one man
in his time plays many parts.
merely players;\nThey have
```

(Este é um bom exemplo do que evitar. Qualquer um que escreva códigos como este deveria ser demitido imediatamente. A ausência de comentários, o código complicado e a brevidade desnecessária tornam difícil a compreensão do que o programa está fazendo. Um programa escrito dessa forma se torna difícil de corrigir e praticamente impossível de manter. É possível que algum dia você suceda alguém em uma tarefa e se depare com um programa como este. Você terá minha compaixão.)

(a) Faça uma cópia deste programa, e da versão original na página 40, de forma que apareçam lado a lado na mesma folha. Onde for possível, relacione as linhas deste programa com as linhas correspondentes do programa mais longo da página 40.

(b) Prepare uma versão do programa conciso com comentários suficientes para explicar melhor o que ele está fazendo (para isso você poderia considerar os comentários do programa original e adaptá-los) e como está fazendo. Não altere nenhuma das instruções executáveis (nem na versão original ou em qualquer outra coisa); apenas adicione comentários.

Weblemas

1.1 Identifique a fonte de todas as citações da peça de Shakespeare no alinhamento do Quadro da página 46.

1.2 Identifique sítios na *web* que forneçam explicações tutoriais *elementares* e/ou demonstrações *on-line* (a) da reação em cadeia da polimerase (PCR – *Polymerase Chain Reaction*), (b) do Southern *blotting*, (c) de mapas de restrição, (d) da árvore de sufixos, (e) do algoritmo *heapsort*. Liste seus endereços eletrônicos (URLs). Escreva um parágrafo com explicações destes termos baseadas nos sítios encontrados.

1.3 A quais filos pertencem as espécies seguintes? (a) Estrela-do-mar, (b) lampreia, (c) tênia, (d) árvore ginkgo, (e) escorpião, (f) água-viva, (g) anêmona-do-mar.

1.4 Quais são os nomes comuns das seguintes espécies? (a) *Acer rubrum*, (b) *Orycteropus afer*, (c) *Beta vulgaris*, (d) *Pyractomena borealis*, (e) *Macrocystis pyrifera*.

1.5 Um café-da-manhã inglês típico consiste em ovos (de galinha) fritos em banha de porco, *bacon*, arenque defumado, cogumelos grelhados, batatas fritas, tomates grelhados, feijão cozido, torradas e chá com leite. Escreva a taxonomia completa dos organismos a partir dos quais estes pratos são derivados.

1.6 Recupere e alinhe as seqüências do citocromo *b* de cavalo, baleia e canguru. (a) Compare o grau de similaridade para cada par de seqüências com o resultado da comparação das seqüências da ribonuclease pancreática destas espécies no Estudo de Caso 1.2. As conclusões tomadas com base na análise das seqüências do citocromo *b* são consistentes com as conclusões tomadas com base na análise da ribonuclease pancreática? (b) Compare a similaridade *relativa* destas seqüências com os resultados da comparação das seqüências da ribonuclease pancreática destas espécies no Estudo de Caso 1.2. As conclusões tomadas com base na análise das seqüências do citocromo *b* são consistentes com as conclusões tomadas com base na análise da ribonuclease pancreática?

1.7 Recupere e alinhe as seqüências da ribonuclease pancreática da baleia Sperm, do cavalo e do hipopótamo. Estes resultados são consistentes com as relações observadas com base no uso dos SINES?

1.8 Observamos que as seqüências de aminoácidos do citocromo *b* de elefantes e de mamute são bastante similares. Uma hipótese para explicar esta observação é que o citocromo *b* funcional *requer* tantos resíduos conservados que os citocromos *b* de todos os animais são similares uns aos outros, assim como as proteínas dos elefantes e mamute. Teste esta hipótese obtendo as seqüências do citocromo *b* de outras espécies de mamíferos, e verifique se as seqüências de aminoácidos do citocromo *b* de espécies pouco relacionadas, ou distantes, são tão similares à seqüência dos elefantes quanto o são as seqüências de elefantes e mamute.

1.9 Recupere e alinhe as seqüências do citocromo *c* de humanos, da cobra cascavel e do lagarto-monitor. Qual par de seqüências parece estar mais proximamente relacionado? Isto o surpreende? Por quê? Ou por que não?

1.10 Envie as seqüências da ribonuclease pancreática de cavalo, baleia-branca e canguru-vermelho (Estudo de Caso 1.2) para o servidor de alinhamentos múltiplos T-coffee: `http://ch.embnet.org/software/TCoffee.html`. O resultado do alinhamento é o mesmo do resultado mostrado no Estudo de Caso 1.2, gerado pelo CLUSTAL-W? Se não for, qual é a diferença?

1.11 Crie um alinhamento múltiplo de seqüências dos *genes* do citocromo *b* mitocondrial dos elefantes africano e indiano e do mamute siberiano. Quantos malpareamentos são observados em cada par de seqüências? Os resultados são consistentes com as conclusões derivadas da comparação de seqüências de proteínas, que afirmam que o mamute está mais proximamente relacionado ao elefante africano do que ao indiano?

1.12 Linnaeus dividiu o reino animal em seis classes: mamíferos, aves, anfíbios (incluindo répteis), peixes, insetos e vermes. Isso implica, por exemplo, que ele considerou crocodilos e salamandras mais proximamente relacionados do que crocodilos e pássaros. Thomas Huxley, por outro lado, no século XIX, reuniu répteis e aves em um mesmo grupo. Para três proteínas apropriadas com homólogas em crocodilos, salamandras e aves, determine a similaridade entre as seqüências homólogas. Quais pares de grupos de animais parecem estar mais proximamente relacionados? Quem estava certo, Linnaeus ou Huxley?

1.13 Quando a última espécie de primata foi descoberta?

1.14 Em quantas novas espécies foram descobertas homólogas da proteína PAX-6 desde que a tabela da página 58 foi compilada?

1.15 Quais são as classificações SCOP para as proteínas da Figura 1.10 (a), (e), (g), (i), (l) e (o)?

1.16 Identifique três proteínas modulares, além da própria fibronectina, que contenham domínios fibronectina III.

1.17 Encontre seis exemplos de doenças, além do diabetes e da hemofilia, que são tratáveis diretamente pela administração da proteína ausente. Em cada caso, qual a proteína administrada?

1.18 Pessoas portadoras de uma variante do gene da apolipoproteína E estão sob risco elevado de qual doença de manifestação tardia? Qual variante oferece o maior risco? O que se sabe sobre o mecanismo pelo qual essas variantes influenciam o desenvolvimento da doença?

1.19 Para aproximadamente 10% dos europeus, o analgésico codeína é ineficaz, pois os pacientes não têm a enzima que converte a codeína na sua molécula ativa, a morfina. Qual é a mutação mais comum que causa esta condição?

1.20 Encontre a página do SCOP encabeçada por "Protein: Thermopin from *Thermobifida fusca*". (a) Qual é o código PDB desta proteína? (b) A qual superfamília esta proteína pertence? (c) Qual homólogo da "Thermopin" de galinha possui estrutura conhecida? (d) Clicando nas conexões disponíveis na página, vá para aquela que apresenta o resumo (*abstract*) do artigo que descreve a determinação da estrutura. Quantos cliques existem no caminho mais curto para chegar a esta página? Quais são as URLs dos sítios intermediários neste caminho?

1.21 Os monotremados, dos quais o exemplo mais conhecido é o ornitorrinco (*Ornithorhynchus anatinus*), formam uma ordem na classe *mammalia*. Apenas algumas espécies de monotremados são conhecidas. (a) Quais destes animais, se houver algum, encontram-se ameaçados de extinção? (b) Onde estes animais são encontrados na natureza? (c) Qual é a natureza da ameaça atual à sua sobrevivência? (d) Quais seqüências gênicas, se houver alguma, de espécies de monotremados ameaçados são conhecidas? (e) Suponha que você queira seqüenciar um gene de uma espécie de monotremado ameaçada. Encontre um zoológico que possua esta espécie.

CAPÍTULO 2

Organização e Evolução de Genomas

Sumário do capítulo

Genomas e proteomas 88
- Genes 89
- Proteomas 91

Espionando a transmissão da informação genética 92
- Mapeamentos entre os mapas 97
- Mapas de alta resolução 98

Selecionando genes em genomas 100

Genomas de procariotos 101
- O genoma da bactéria *Escherichia coli* 102
- O genoma da arqueobactéria *Methanococcus jannaschii* 105
- O genoma de um dos organismos mais simples: *Mycoplasma genitalium* 106

Genomas de eucariotos 107
- O genoma de *Saccharomyces cerevisiae* (fermento de padeiro) 109
- O genoma de *Caenorhabditis elegans* 112
- O genoma de *Drosophila melanogaster* 114
- O genoma de *Arabidopsis thaliana* 115

O genoma do *Homo sapiens* (o genoma humano) 116
- Genes que codificam proteínas 117
- Seqüências de repetições 119
- RNA 120

Polimorfismos de um único nucleotídeo (SNPs) 121

Diversidade genética na antropologia 122
- Diversidade genética e identificação pessoal 123
- Análise genética da domesticação do gado 124

Evolução de genomas 124
- Por gentileza, passem os genes: transmissão horizontal de genes 128
- Genômica comparativa de eucariotos 129

Leituras recomendadas 131

Exercícios, problemas e weblemas 132

Objetivos do aprendizado

1. Conhecer o tamanho básico e os princípios da organização de genomas simples e complexos.
2. Entender os conceitos de proteômica – os padrões de expressão de proteínas – e as técnicas utilizadas para estudá-los, tal como os microarranjos.
3. Saber como genomas são analisados e conhecer as relações entre seqüências de genes e características fenotípicas, incluindo doenças hereditárias.
4. Compreender a idéia geral dos conteúdos de genomas específicos e as diferenças sistemáticas entre os genomas de procariotos e eucariotos.
5. Reconhecer a importância e a dificuldade de, a partir da seqüência completa de um genoma, derivar as seqüências de aminoácidos das proteínas codificadas e atribuir funções às mesmas.
6. Perceber que os genomas publicados registram as características de apenas um único indivíduo; que existe uma variação considerável dentro de uma população de indivíduos; e que existe uma variação ainda maior entre populações distintas de organismos de uma mesma espécie.
7. Apreciar o poder da genômica comparativa na identificação das características responsáveis pelas diferenças entre as espécies – por exemplo, o que nos torna humanos?

Genomas e proteomas

O genoma de uma bactéria típica se apresenta como uma única molécula de DNA que, se estendida, mediria cerca de 2 mm de comprimento. (A própria célula tem um diâmetro de cerca de 0,001 mm.) O DNA de organismos superiores está organizado em cromossomos – células humanas normais contêm 23 pares de cromossomos. A quantidade total de informação genética por célula – a seqüência de nucleotídeos de DNA – é aproximadamente constante para todos os membros de uma espécie, mas varia bastante entre espécies (ver Quadro para uma lista maior de exemplos):

Organismo	Tamanho do genoma (pares de bases)
Vírus Epstein-Barr	$0{,}172 \times 10^6$
Bactéria (*E. coli*)	$4{,}6 \times 10^6$
Levedura (*S. cerevisiae*)	$12{,}5 \times 10^6$
Verme nematódeo (*C. elegans*)	$100{,}3 \times 10^6$
Agrião Thale (*A. thaliana*)	$115{,}4 \times 10^6$
Mosca-das-frutas (*D. melanogaster*)	$128{,}3 \times 10^6$
Humanos (*H. sapiens*)	3.223×10^6

Nem todo o DNA codifica proteínas. Ademais, alguns genes existem em múltiplas cópias. Portanto, a quantidade de informação sobre seqüências de proteínas em uma célula não pode ser facilmente estimada a partir do tamanho do genoma.

Genes

Um gene que codifica uma determinada proteína corresponde a uma seqüência de nucleotídeos ao longo de uma ou mais regiões de uma molécula de DNA. A seqüência de DNA é colinear à seqüência protéica codificada. Nas espécies em que o material genético é o DNA de dupla hélice, os genes podem aparecer em ambas as fitas. Genes de bactérias são regiões contínuas de DNA. Assim, a unidade funcional de informação da seqüência genética de uma bactéria é uma cadeia de $3N$ nucleotídeos codificando uma cadeia de N aminoácidos, ou uma cadeia de N nucleotídeos que codifica uma molécula de RNA estrutural com N resíduos. Esta cadeia ou seqüência de nucleotídeos ou aminoácidos, incluindo suas anotações, forma uma entrada típica em um dos bancos de dados de seqüências genéticas.

Tamanhos de alguns genomas

Organismo	Número de pares de bases	Número de genes	Comentário
ΦX-174	5.386	10	vírus que infecta *E. coli*
Mitocôndria humana	16.569	37	organela subcelular
Vírus Epstein-Barr (EBV)	172.282	80	causador da mononucleose
Mycoplasma pneumoniae	816.394	680	causador de epidemias de pneumonia cíclica
Rickettsia prowazekii	1.111.523	834	bactéria causadora de epidemias de tifo
Treponema pallidum	1.138.011	1.039	bactéria causadora da sífilis
Borrelia burgdorferi	1.471.725	1.738	bactéria causadora da doença de Lyme
Aquifex aeolicus	1.551.335	1.749	bactéria encontrada em fontes termais
Thermoplasma acidophillum	1.564.905	1.509	arqueobactéria sem parede celular
Campylobacter jejuni	1.641.481	1.708	causa freqüente de envenenamentos alimentares
Helicobacter pylori	1.667.867	1.589	principal causa de úlcera estomacal
Methanococcus jannaschii	1.664.970	1.783	arqueobactéria termófila
Haemophilus influenzae	1.830.138	1.738	bactéria causadora de infecções do ouvido médio
Thermotoga maritima	1.860.725	1.879	bactéria marinha
Archaeoglobus fulgidus	2.178.400	2.437	outra arqueobactéria
Deinococcus radiodurans	3.284.156	3.187	bactéria resistente à radiação
Synechocystis	3.573.470	4.003	cianobactéria ou "alga azul"

→

Tamanhos de alguns genomas (*continuação*)

Vibrio cholerae	4.033.460	3.890	causa da cólera
Mycobacterium tuberculosis	4.411.532	3.959	causa da tuberculose
Bacillus subtilis	4.214.814	4.779	bactéria popular na biologia molecular
Escherichia coli	4.639.221	4.377	organismo predileto dos biólogos moleculares
Pseudomonas aeruginosa	6.264.403	5.570	um dos maiores genomas procarióticos já seqüenciados
Saccharomyces cerevisiae	12.495.682	5.770	levedura, primeiro organismo eucariótico seqüenciado
Caenorhabditis elegans	103.006.709	20.598	verme
Arabidopsis thaliana	115.409.949	25.498	planta florífera (angiosperma)
Drosophila melanogaster	128.343.463	13.525	mosca-das-frutas
Takifugu rubripes	329×10^6	34.080?	peixe fugo (baiacu japonês)
Humanos	3.223×10^6	23.000?	
Trigo	16×10^9	30.000	
Salamandra	10^{11}	?	
Psilotum nudum	$2{,}5 \times 10^{11}$	?	feto – planta simples

Em eucariotos, as seqüências de nucleotídeos que codificam as seqüências de aminoácidos de proteínas individuais estão organizadas de uma maneira mais complexa. Freqüentemente, um gene se apresenta dividido em segmentos separados no DNA genômico. Um **éxon** (região expressa) é um segmento de DNA mantido no RNA mensageiro maduro que o ribossomo traduz em proteína. Um **íntron** é um fragmento de DNA interveniente entre dois éxons. A maquinaria celular une (*splices*) os segmentos apropriados nos transcritos de RNA, utilizando como base seqüências sinalizadoras flanqueando os éxons. Muitos íntrons são bastante longos – em alguns casos, consideravelmente mais longos do que os próprios éxons.

O controle da informação organiza a expressão de genes. Mecanismos de controle podem ligar ou desligar genes (ou mediar a regulação ótima da expressão de um gene) em resposta à concentração de nutrientes, ou ao estresse, ou desencadear programas complexos de desenvolvimento durante a vida de um organismo. Muitas regiões de controle do DNA estão localizadas próximas a segmentos de DNA que codificam proteínas. Elas contêm **seqüências sinalizadoras** que servem de sítio de ligação para as moléculas que transcrevem a seqüência de DNA, ou seqüências que ligam moléculas reguladoras que podem **bloquear** a transcrição. Os genomas de bactérias contêm exemplos de genes contíguos codificando diversas proteínas que catalisam etapas sucessivas em uma seqüência de reações integradas, estando todos os genes sob o controle da mesma seqüência reguladora. F. Jacob, J. Monod e E. Wollman chamaram este arranjo de **óperons**. Pode-se entender facilmente a utilidade de um mecanismo

de controle paralelo. Em animais, a metilação do DNA fornece os sinais para a expressão tecido-específica de genes regulados durante o desenvolvimento.

Os produtos de certos genes podem levar a célula a cometer suicídio – um processo denominado **apoptose**. Defeitos nos mecanismos de apoptose podem levar ao crescimento descontrolado de células, como observado em alguns tipos de cânceres, e a estimulação desses mecanismos é uma das propostas atuais para a terapia do câncer.

A conclusão a que se pode chegar é que reduzir os dados genéticos às seqüências codificadoras individuais é ocultar a natureza bastante complexa das inter-relações entre eles e ignorar os aspectos históricos e integrativos do genoma. Robbins caracterizou esta situação:

"...Considere que os 3,2 gigabytes do genoma humano são equivalentes a 3,2 gigabytes de arquivos armazenados no disco rígido de algum sistema de computador de arquitetura desconhecida. A obtenção da seqüência é equivalente à visualização dos conteúdos armazenados nesse disco rígido. Entender a seqüência corresponde a desvendar o sistema de computador desconhecido (tanto o *hardware* quanto os 3,2 gigabytes de *software*) percorrendo o caminho inverso, até um conjunto completo de especificações de projeto e manutenção."

...

"A engenharia reversa da seqüência é complicada pelo fato de que a imagem resultante do disco rígido não será uma cópia arquivo por arquivo, mas sim um amontoado de bytes na ordem em que foram registrados no computador. Além disso, sabe-se que os arquivos estão fragmentados. Ademais, parte do disco rígido contém arquivos apagados ou outras informações não importantes (lixo). Uma vez que esse lixo tenha sido identificado e descartado, e os arquivos fragmentados reunidos, a engenharia reversa dos códigos só poderá ser realizada com o conhecimento parcial, e às vezes incorreto, da CPU em que o código executa. De fato, a dedução da estrutura e da função da CPU é parte do projeto, uma vez que parte dos 3,2 gigabytes são as especificações binárias para o processo de fabricação-assistida-por-computador que manufatura a CPU. Além disso, deve-se ainda considerar que o enorme banco de dados também contém códigos gerados do resultado de, literalmente, milhares de revisões de manutenção realizadas pelo pior conjunto possível de ciberpiratas oportunistas que se deleitam em tramóias espertas, como escrever códigos que se automodificam e que são dependentes dos caprichos de sistemas maldocumentados."

Robbins, R. J. Challenges in the Human Genome Project. *IEEE Engineering in Medicine and Biology*, 11:25-34, (1992) (© 1992 IEEE).

Proteomas

O **projeto proteoma** é um programa em larga escala que lida de forma integral com os padrões de expressão de proteínas em sistemas biológicos, de modo complementar e suplementar aos projetos genoma.

Que tipos de dados gostaríamos de avaliar e quais técnicas experimentais bem estabelecidas existem para determiná-los? O objetivo principal é a descrição espaço-temporal da distribuição de proteínas no organismo. As taxas de síntese de diferentes proteínas variam entre diferentes tecidos e tipos celulares, bem como com o estado da atividade celular. Existem métodos disponíveis para a análise de padrões de transcrição de genes. Porém, como as proteínas são "recicladas" em velocidades distintas, é também necessário quantificá-las diretamente. A eletroforese bidimensional em gel de poliacrilamida (2D PAGE) de alta resolução mostra o padrão do conteúdo de proteínas em uma amostra. As técnicas de espectroscopia de massa identificam as proteínas separadas a partir da amostra. Retornaremos a esses tópicos no Capítulo 6.

Em princípio, um banco de dados de seqüências de aminoácidos de proteínas é inerente ao banco de dados de seqüências de nucleotídeos de DNA em virtude do código genético. Na verdade, novos dados de seqüências de proteínas estão sendo agora determinados quase exclusivamente pela tradução de seqüências de DNA, em vez do seqüenciamento direto de proteínas. (Historicamente, o problema químico de determinação direta da seqüência de aminoácidos de proteínas foi resolvido antes do estabelecimento do código genético e antes do desenvolvimento dos métodos para a determinação da seqüência de nucleotídeos do DNA. O seqüenciamento da insulina em 1955 por F. Sanger provou que as proteínas tinham uma seqüência definida de aminoácidos, uma proposição até então hipotética.)

Deveria ser feita alguma distinção entre as seqüências de aminoácidos determinadas diretamente a partir do seqüenciamento de proteínas e aquelas determinadas pela tradução do DNA? Primeiro, devemos assumir que é possível identificar corretamente, dentro de uma torrente de dados de DNA, as regiões que codificam proteínas. Os programas de reconhecimento de padrões que tratam da resolução deste problema estão sujeitos a três tipos de erros: uma seqüência de proteína genuína pode ser totalmente ignorada, uma proteína incompleta pode ser descrita, ou um gene pode ser recomposto (*spliced*) incorretamente. Algumas variações deste tema aumentam ainda mais a complexidade do problema: os genes de diferentes proteínas podem se sobrepor, ou os genes podem ser formados de éxons em diferentes combinações e em tecidos distintos. De modo inverso, algumas seqüências genéticas que parecem codificar proteínas podem, de fato, ser defeituosas ou não ser expressas. *Uma proteína inferida a partir da seqüência de um genoma é apenas um objeto hipotético até que experimentos comprovem a sua existência.*

Segundo, em muitos casos, a expressão de um gene produz uma molécula que deve ser modificada no interior da célula para gerar uma proteína *madura* que difere significativamente daquela sugerida pela tradução da seqüência do gene. Em muitos casos, os detalhes omitidos das **modificações pós-traducionais** – o análogo molecular de *body piercing* – são bastante importantes. Modificações pós-traducionais incluem adição de ligantes (por exemplo, o grupo heme ligado covalentemente ao citocromo *c*), glicosilação, metilação, excisão de peptídeos, entre muitos outros. Padrões de pontes dissulfeto – ligações químicas fundamentais entre resíduos de cisteínas – não podem ser deduzidos pela seqüência de aminoácidos. Em alguns casos, o mRNA é editado antes da sua tradução, gerando alterações nas seqüências de aminoácidos que não são dedutíveis das seqüências dos genes.

Espionando a transmissão da informação genética

Como a informação hereditária é armazenada, transmitida e implementada talvez seja o problema fundamental da biologia. Existem três tipos essenciais de mapas (ver Quadro):

1. Mapas de ligações de genes
2. Padrões de bandeamento de cromossomos
3. Seqüências de DNA

Esses mapas representam tipos bastante diferentes de dados. Genes, conforme descobertos por Mendel, eram entidades totalmente abstratas. Os cromossomos são objetos físicos, os padrões de bandeamento, seus pontos de referên-

Mapas de genes, mapas de cromossomos e mapas de seqüências

1. Um mapa de genes é determinado classicamente pela observação de padrões de hereditariedade. Os grupos de ligação e as freqüências de recombinação podem detectar quando os genes estão em um mesmo cromossomo ou em cromossomos distintos e, para genes em um mesmo cromossomo, a distância entre eles também pode ser determinada. Esse tipo de mapa se baseia no princípio de que quanto mais distante estiverem dois genes ligados, maior será a probabilidade de ocorrer recombinação entre eles, através de *crossing over* durante a meiose. De fato, dois genes em um mesmo cromossomo, mas bastante distantes um do outro, parecerão não estar ligados. A unidade de comprimento no mapa de genes é o Morgan, definido pela relação na qual 1 cM corresponde a uma freqüência de recombinação de 1%. (Agora sabemos que 1 cM é ~1 $\times 10^6$ pb em humanos, mas ele varia conforme a localização no genoma e com a distância entre os genes.)
2. Mapas de padrões de bandeamento de cromossomos Os cromossomos são objetos físicos. Padrões de bandas são suas características visíveis. Sua nomenclatura é determinada da seguinte forma: em muitos organismos, os cromossomos são numerados por ordem de tamanho, sendo o número 1 o maior. Os dois braços de cromossomos humanos, separados pelo centrômero, são chamados de braço p (do adjetivo francês *petite* = pequeno) e braço q (*q* segue *p* no alfabeto latino). Regiões do cromossomo são numeradas p1, p2, ... e q1, q2... para fora, a partir do centrômero. Dígitos subseqüentes indicam as subdivisões das bandas. Por exemplo, algumas bandas no braço q do cromossomo 15 humano são denominadas 15q11.1, 15q11.2, 15q12. As bandas 15q11 e 15q12 foram definidas inicialmente, e a banda 15q11 foi, posteriormente, dividida em 15q11.1 e 15q11.2. Deleções nesta região do genoma estão associadas com as síndromes de Prader-Willi e de Angleman. Essas síndromes possuem uma característica interessante: suas conseqüências clínicas dependem da origem do cromossomo afetado, se é herança paterna ou materna. Esta observação da impressão genômica mostra que a informação genética em um ovo ou óvulo fertilizado não é simplesmente o conjunto de seqüências de DNA puro oriundo dos pais. Cromossomos de origem paterna e materna apresentam estados distintos de metilação, sinais para a expressão diferenciada de seus genes. O processo de modificação do DNA, que ocorre na diferenciação, durante o desenvolvimento, ainda está presente no zigoto.
3. A seqüência de DNA Fisicamente, constitui-se da seqüência de nucleotídeos em uma molécula; computacionalmente, é um conjunto de caracteres: A, T, G e C. Os genes são regiões da seqüência, em muitos casos, interrompida por regiões não-codificadoras.

cia visíveis. Apenas com as seqüências de DNA estamos lidando diretamente com a informação hereditária armazenada em sua forma física.

Fazer conexões entre esses três tipos de dados foi a grande conquista da biologia do século passado. Os primeiros passos – e foram realmente passos gigantes – provaram que, para qualquer cromossomo, os mapas são arranjos unidi-

mensionais e que, de fato, eles são colineares. Qualquer criança que freqüenta a escola sabe, hoje, que os genes estão dispostos ao longo dos cromossomos e que cada gene corresponde a uma seqüência de DNA. Não obstante, as comprovações destas declarações renderam alguns prêmios Nobel.

A partição de uma longa molécula de DNA – por exemplo, o DNA de um cromossomo inteiro – em fragmentos de tamanho conveniente para clonagem e seqüenciamento exige mapas adicionais que apresentem a ordem dos fragmentos de tal forma que a seqüência íntegra possa ser reconstruída a partir das seqüências dos fragmentos. Uma endonuclease de restrição é uma enzima que cliva o DNA em seqüências específicas, geralmente com cerca de 6 pb. Clivando o DNA com diversas enzimas de restrição com especificidades diferentes, obtém-se um conjunto de fragmentos com sobreposições. Com base no tamanho dos fragmentos, é possível construir um **mapa de restrição**, estabelecendo a ordem e a distância entre os sítios de clivagem das enzimas de restrição. Uma mutação em um desses sítios de clivagem mudará os tamanhos dos fragmentos produzidos pela enzima correspondente, permitindo que a mutação seja localizada no mapa.

As enzimas de restrição podem produzir fragmentos razoavelmente grandes de DNA. A clivagem do DNA em fragmentos menores, que são clonados e ordenados pelas sobreposições nas seqüências (como o exemplo utilizando textos nas páginas 39 e 40), produz uma análise mais detalhada do DNA denominada **mapa de contíguos**.

No passado, as relações entre cromossomos, genes e seqüências de DNA foram essenciais na identificação das deficiências moleculares responsáveis pelas doenças hereditárias, como a doença de Huntington ou a fibrose cística. O seqüenciamento do genoma humano modificou esta situação radicalmente.

Dada uma doença ocasionada por uma proteína defeituosa:

- Se soubermos qual é a proteína envolvida, podemos desenvolver terapias por meio de métodos racionais.
- Se soubermos qual é o gene envolvido, podemos desenvolver testes para identificar indivíduos doentes e portadores.
- Em muitos casos, o conhecimento da localização cromossômica do gene não é necessário para a terapia nem para a detecção; é necessário apenas para a identificação do gene, criando uma ponte entre os padrões de hereditariedade e a seqüência de DNA. (Isso não é verdadeiro para doenças originadas de anormalidades cromossômicas.)

Por exemplo, no caso da anemia falciforme, sabemos qual é a proteína envolvida. A doença é causada por uma única mutação na hemoglobina. Podemos prosseguir diretamente para o planejamento de fármacos. Precisamos da seqüência de DNA apenas para testes genéticos e aconselhamento de pacientes. Ao contrário, se nem a proteína ou o gene fossem conhecidos, precisaríamos traçar, de alguma forma, uma relação entre o fenótipo e o gene, um processo chamado de **clonagem posicional** ou **genética reversa**. A clonagem posicional costumava envolver um tipo de cascata de relações entre o mapa de genes, o mapa de cromossomos e a seqüência de DNA. Veremos mais tarde que desenvolvimentos recentes nesta área simplificaram este processo.

Padrões de herança identificam o tipo de defeito genético responsável por uma doença ou condição. Eles mostram, por exemplo, que a doença de Huntington e a fibrose cística são causadas por apenas um gene cada. Para identificar o gene responsável pela fibrose cística, foi necessário começar com o mapa de genes, utilizando padrões de ligação de hereditariedade em famílias afetadas a fim de

se localizar o gene afetado em uma região particular de um cromossomo específico. Conhecendo a região principal do cromossomo, foi então possível pesquisar o DNA daquela região para identificar genes candidatos e, finalmente, localizar o gene responsável e seqüenciá-lo (ver Quadros Identificação do gene da fibrose cística e Clonagem posicional: encontrando o gene da fibrose cística). Ao contrário, muitas doenças não apresentam herança simples ou, mesmo se apenas um único gene está envolvido, a hereditariedade cria somente uma predisposição, cujas conseqüências clínicas dependem de fatores ambientais. A seqüência completa do genoma humano e a determinação de padrões de expressão serão essenciais para identificar os componentes genéticos desses casos mais complexos.

Identificação do gene da fibrose cística

A fibrose cística, uma doença conhecida pelo folclore desde, pelo menos, a Idade Média e, pela ciência, há cerca de 500 anos, é uma doença hereditária autossômica recessiva. Seus sintomas incluem obstrução intestinal, fertilidade reduzida, incluindo anomalias anatômicas (especialmente em homens), e obstrução e infecção recorrentes dos pulmões – a causa principal de morte, agora que estão disponíveis tratamentos eficazes para os sintomas gastrintestinais. Aproximadamente metade dos afetados morre antes dos 32 anos, e poucos sobrevivem além dos 50 anos. A fibrose cística afeta uma em cada 2.500 pessoas nas populações americana e européia. Aproximadamente um em cada 25 brancos são portadores do gene mutante, e um em cada 65 afro-americanos. A proteína defeituosa na fibrose cística também atua como um receptor para absorção de *Salmonella typhi*, o patógeno causador da febre tifóide. A maior resistência à febre tifóide dos heterozigotos – que não desenvolvem a fibrose cística, mas são portadores do gene mutante – explica por que o gene não foi eliminado da população.

O padrão de herança mostrou que a fibrose cística era o efeito de um único gene. Entretanto, a proteína envolvida não era conhecida. Ela foi descoberta *por meio* do gene.

Observações clínicas forneceram pistas úteis aos caçadores de genes (pesquisadores). Sabia-se que o problema estava relacionado com o transporte de íons Cl^- nos tecidos epiteliais. O conhecimento popular há muito tempo reconhecia que crianças com excesso de sal no suor – fato percebido ao se beijar a testa das crianças – não viviam por muitos anos. Estudos modernos de fisiologia mostraram que os tecidos epiteliais de pacientes com fibrose cística não conseguem reabsorver cloreto. Quando os pesquisadores estavam próximos da descoberta do gene, a distribuição esperada de sua expressão nos tecidos e o tipo de proteína implicada foram guias bastante úteis.

Em 1989, o gene da fibrose cística foi isolado e seqüenciado. Este gene – chamado de CFTR (*cystic fibrosis transmembrane conductance regulator*) – codifica uma proteína de 1.480 aminoácidos que normalmente forma um canal epitelial de Cl^- regulado por AMP cíclico. O gene, composto de 24 éxons, estende-se por 250 kpb. Para 70% dos alelos mutantes, a mutação observada é a deleção de três pares de base, que resulta na deleção do resíduo 508Phe da proteína. Essa mutação é indicada por del508. O efeito dessa

→

Identificação do gene da fibrose cística (*continuação*)

deleção é a translocação defectiva da proteína, que é degradada no retículo endoplasmático em vez de ser transportada para a membrana celular.

O teste *in utero* para a fibrose cística se baseia na coleta de DNA fetal. Um oligonucleotídeo iniciador *de* PCR é desenhado para originar um produto de 154 pb a partir do alelo normal e um produto de 151 pb a partir do alelo mutante del508.

Os médicos aproveitaram o fato de que os tecidos afetados das vias aéreas são de fácil acesso para desenvolver terapias gênicas. Os adenovírus geneticamente modificados e aerossolizados dentro das vias respiratórias podem distribuir o gene correto nos tecidos epiteliais.

Clonagem posicional: encontrando o gene da fibrose cística

O processo pelo qual o gene foi localizado é chamado de *clonagem posicional* ou *genética reversa*.

- Um levantamento nas famílias afetadas por marcadores relacionados mostrou que o gene da fibrose cística estava próximo a uma repetição em *tandem* de número variável (VNTR – *variable number tandem repeat*) já conhecida, a DOCR-917. Híbridos de células somáticas localizaram essa repetição no cromossomo 7, banda q3.
- Outros marcadores encontrados foram mais fortemente relacionados ao gene-alvo. Ele estava entre a repetição VNTR do oncogene MET e uma segunda VNTR, a D7S8. O gene-alvo estava a 1,3 cM do gene MET e a 0,9 cM de D7S8 – localizando-o em uma região de aproximadamente 1 a 2 milhões de pares de base. Uma região deste tamanho poderia conter, sem problemas, entre 100 e 200 genes.
- Os padrões de herança de marcadores adicionais dentro desta região limitaram a posição do gene com uma precisão de 500 kpb. Uma técnica chamada *jumping* de cromossomos tornou a exploração da região mais eficaz.
- Uma região de 300 kpb a uma distância adequada dos marcadores foi clonada. Amostras de DNA foram isoladas dessa região para procurar por genes ativos caracterizados pela seqüência 5'-CCGG. (A endonuclease de restrição HpaII foi utilizada nesta etapa; ela cliva este DNA somente quando o segundo C da seqüência não está metilado, ou seja, quando o gene está ativo.)
- A identificação dos genes nesta região foi realizada por seqüenciamento.
- A procura, em animais, por genes similares aos genes esperados revelou quatro candidatos promissores. A checagem desses candidatos, contra uma biblioteca de cDNA de glândulas sudoríparas de pacientes com fibrose cística e controles saudáveis, identificou uma amostra com a distribuição tecidual correta para o padrão de expressão esperado do gene responsável pela fibrose cística. Um segmento codificador longo apresentava as propriedades corretas e, de fato, correspondia ao éxon do gene da fibrose cística. A maioria dos pacientes com fibrose cística possui uma alteração comum na seqüência desse gene – uma deleção de três pares de base, resultando na deleção do resíduo 508Phe da proteína.

Provas de que o gene foi identificado corretamente incluíam:

- 70% dos alelos da fibrose cística apresentam a deleção. Ela não é observada em pessoas que não sofrem da doença nem nos prováveis portadores.
- A expressão do gene tipo selvagem em células isoladas de pacientes com a doença restaura o transporte normal de cloreto.
- A inativação do gene homólogo em camundongos induz o fenótipo da fibrose cística.
- O padrão de expressão gênica ocorre nos órgãos onde é esperado.
- A proteína codificada pelo gene conteria um domínio transmembrana, consistente com a atividade de transporte.

Mapeamentos entre os mapas

Um mapa de ligação gênica pode ser calibrado com os padrões de bandeamento de cromossomos por meio da observação de indivíduos com deleções ou translocações de partes de cromossomos. Os genes responsáveis pelas alterações fenotípicas associadas com uma deleção devem necessariamente estar na porção removida. As translocações estão relacionadas com padrões alterados de ligação e recombinação.

Existem diversas técnicas para ajustar os padrões de bandeamento de cromossomos com seqüências de DNA de genes individuais:

- Na hibridização *in situ* por fluorescência (FISH), a seqüência utilizada como sonda é marcada com corantes fluorescentes. A sonda é hibridizada com os cromossomos, e a localização cromossômica da sonda ligada aparece claramente em uma fotografia (ver Gravura II). A resolução típica é de ~10^5 pares de base, mas técnicas novas mais especializadas atingem resoluções maiores de até 1 kpb. A realização de FISH simultâneas com duas sondas pode detectar a ligação e ainda estimar as distâncias genéticas. Isso é importante nas espécies em que o tempo de geração é longo o suficiente para tornar as técnicas genéticas padrões inconvenientes. A FISH também pode ser utilizada para detectar anomalias cromossômicas.
- **Híbridos de células somáticas** são células de roedores contendo poucos, um ou até mesmo uma seqüência parcial de cromossomos humanos. (Os fragmentos de cromossomos são gerados pela irradiação de células humanas antes da fusão. Estas linhagens são chamadas de **híbridos de radiação**.) A hibridização de uma seqüência sonda com uma amostra de híbridos de células somáticas, detectada por fluorescência, identifica o cromossomo que contém a sonda. Este método foi substituído pelo uso de clones de levedura, bactérias ou mesmo de fagos contendo fragmentos de DNA humano em cromossomos artificiais (YACs, BACs e PACs).

Certamente, para genomas cujas seqüências foram determinadas por completo, estes métodos se tornaram obsoletos. Dada uma seqüência de DNA, basta procurá-la no genoma (but mind the gaps*).

* N. de T. A expressão em inglês "mind the gaps" refere-se ao espaço que há entre a plataforma e o vagão do trem no metrô de Londres, Inglaterra. No alto-falante das estações de metrô, esta frase é constantemente repetida para alertar os passageiros. Muitos já ficaram com os pés ou as pernas presos nestas aberturas. No contexto deste parágrafo, portanto, a frase quer lembrar o leitor das lacunas que podem existir nos genomas por não estarem 100% completos.

Mapas de alta resolução

Antigamente, os genes eram as únicas partes visíveis de genomas. Hoje, os marcadores não se limitam aos genes com efeitos fenotípicos observáveis, os quais são, de qualquer forma, muito esparsos para a elaboração adequada de um mapa de alta resolução do genoma humano. Agora que podemos consultar diretamente a seqüência de DNA, qualquer característica deste que varie entre indivíduos pode ser utilizada como marcador, incluindo:

- repetições em *tandem* de número variável (VNTRs), são também chamadas de minissatélites. As VNTRs contêm regiões de 10 a 100 pb repetidas um número variável de vezes – a mesma seqüência, um número diferente de repetições. Em qualquer indivíduo, as VNTRs baseadas no mesmo motivo de repetição podem aparecer apenas uma vez no genoma; ou várias vezes, com tamanhos diferentes em cromossomos diferentes. A distribuição dos tamanhos das repetições é o marcador. A herdabilidade de VNTRs pode ser acompanhada em uma família e mapeada a um fenótipo de doença, como qualquer outra característica. As VNTRs foram os primeiros dados de seqüências genéticas utilizados na identificação pessoal – impressão digital genética – em casos criminais e de paternidade.

  No início, as VNTRs eram observadas por meio da produção de polimorfismos de comprimento de fragmentos de restrição (RFLPs) a partir delas próprias. As VNTRs são geralmente flanqueadas por sítios de reconhecimento da mesma enzima de restrição que irá clivá-las. Os resultados podem ser aplicados em um gel e detectados por Southern *blotting*. Observe a diferença: as VNTRs são características de seqüências genômicas; os RFLPs são misturas artificiais de pequenos fragmentos de DNA, gerados em laboratório, para a identificação das VNTRs.

  É muito mais fácil e mais eficiente medir os tamanhos das VNTRs por meio de suas amplificações por PCR, e este método substituiu o uso das enzimas de restrição.

- polimorfismos de repetições curtas em *tandem* (STRPs), também chamados de microssatélites, são regiões de apenas 2 a 5 pb repetidos muitas vezes; tipicamente de 10 a 30 cópias consecutivas. Eles possuem diversas vantagens como marcadores sobre as VNTRs, uma das quais é a sua distribuição mais uniforme ao longo do genoma humano.

Não existe razão para que estes marcadores estejam inseridos em genes que são expressos e, geralmente, não estão. (As repetições CAG no gene da huntingtina e em outros genes relacionados com doenças são exceções.)

Distribuições de marcadores microssatélites simplificam substancialmente a identificação de genes. É interessante comparar um projeto recente para identificar um gene envolvido em uma doença, agora que a seqüência do genoma humano está disponível, com os estudos clássicos, tal como a identificação do gene da fibrose cística (ver Quadro Identificação de um gene da síndrome de Berardinelli-Seip).

Outras técnicas de mapeamento lidam mais diretamente com a seqüência de DNA, podendo acelerar o processo de identificação de genes:

- Um contíguo, ou mapa de clones contíguos, é uma série de clones de DNA, com seqüências que apresentam sobreposições de ordenamento conhecido, ao longo de um cromossomo de um organismo de interesse – por exemplo, o humano – armazenados em células de levedura como YACs (*Yeast Artificial Chromossomes*, ou Cromossomos Artificiais de Levedura) ou em

Identificação de um gene da síndrome de Berardinelli-Seip

A síndrome de Berardinelli-Seip (lipodistrofia generalizada congênita) é uma doença autossômica recessiva. Seus sintomas incluem ausência de gordura corporal, diabetes resistente à insulina e taxa elevada de crescimento esquelético.

Para determinar o gene envolvido, o grupo liderado por J. Magré submeteu o DNA de membros de famílias afetadas à análise de ligação de genes e ao mapeamento de homozigose com bancos que cobriam todo o genoma, com cerca de 400 marcadores microssatélites de localização genética conhecida, com espaçamento médio de 10 cM. Neste procedimento, uma distribuição fixa de oligonucleotídeos iniciadores específicos para a amplificação e análise de cada marcador foi utilizada para comparar o DNA total de indivíduos afetados com o DNA total de seus parentes saudáveis. As análises revelaram os comprimentos das repetições associadas a cada microssatélite. Para cada microssatélite, cada comprimento observado é um alelo. A identificação dos marcadores microssatélites mais proximamente relacionados com o fenótipo localiza o gene desejado. As medidas foram realizadas de forma eficiente e em paralelo utilizando conjuntos de oligonucleotídeos iniciadores e equipamentos comerciais.

Dois marcadores no cromossomo 11q13 – D11S4191 e D11S987 – segregavam com a doença, e alguns indivíduos afetados, nascidos em famílias consangüíneas, eram homozigotos para eles. Uma investigação mais detalhada, mapeando com marcadores adicionais, localizou o gene no cromossomo 11 em uma região com cerca de 2,5 Mpb.

Existem 27 genes nessa região e em sua vizinhança. O seqüenciamento desses genes em um grupo de pacientes identificou uma deleção de três éxons em um deles. Foi provado que este era o gene da doença, pela comparação de suas seqüências em membros das famílias estudadas e pela demonstração de uma correlação entre a presença da síndrome e anomalias no gene. Nenhum dos outros 26 genes na região identificada como suspeita apresentou tais alterações correlacionadas.

Estudos anteriores haviam identificado um gene diferente, *BSCL1*, no cromossomo 9q34, em outras famílias portadoras da mesma síndrome. O gene *BSCL1* ainda não foi caracterizado. É possível que anomalias nesses dois genes produzam o mesmo efeito, pois seus produtos fazem parte de uma via comum que pode ser bloqueada pela disfunção de um ou outro.

O gene no cromossomo 11, *BSCL2*, contém 11 éxons, distribuídos em mais de 14 kpb. Ele codifica uma proteína de 398 resíduos de aminoácidos denominada seipina. As alterações já observadas nesse gene incluem deleções grandes e pequenas e substituições de um único aminoácido. Os efeitos dessas alterações são consistentes com a perda da funcionalidade da proteína, seja pela mudança da fase de leitura ou pelo truncamento da proteína, ou ainda pela mutação não-sinônima Ala212→Pro que seguramente interfere com a estabilidade de uma hélice ou folha β na estrutura.

A seipina possui homólogas em camundongos e em *Drosophila*. Não há indícios da função de nenhuma de suas homólogas, embora tenha sido predita a presença de domínios transmembrana formados por hélices. O que sugere algumas explicações sobre a etiologia de alguns aspectos da síndrome é o padrão de expressão, que é maior no cérebro e nos testículos. Isso pode ser consistente com estudos endocrinológicos anteriores da síndrome de

→

Identificação de um gene da síndrome de Berardinelli-Seip (*continuação*)

Berardinelli-Seip que identificaram um problema na regulação da liberação de hormônios pituitários pelo hipotálamo. A descoberta de uma proteína de função desconhecida, envolvida na síndrome, abre o caminho para a investigação do que pode ser, provavelmente, uma nova via biológica.

bactérias como BACs (*Bacterial Artificial Chromossomes*, ou Cromossomos Artificiais de Bactérias). Um mapa de contíguos é capaz de produzir um mapeamento refinado de um genoma. Em um YAC, o DNA humano é integrado de forma estável em um pequeno cromossomo extra em uma célula de levedura. Um YAC pode conter até 10^6 pares de base. Em princípio, todo o genoma humano poderia ser representado em 10.000 clones YAC. Em um BAC, o DNA humano é inserido em um plasmídeo em uma célula de *E. coli*. (Um plasmídeo é um pequeno segmento de DNA de dupla fita presente em adição ao genoma principal, geralmente, mas nem sempre, na forma circular.) Um BAC pode tolerar cerca de 250.000 pb. Apesar da sua capacidade reduzida, os BACs são preferidos aos YACs, por sua maior estabilidade e fácil manuseio.

- Um **sítio com seqüência marcada** (STS) é uma região pequena e seqüenciada do DNA, tipicamente de 200 a 600 pb, que aparece em uma posição única do genoma. Ela não precisa ser polimórfica. Um STS pode ser mapeado no genoma pela reação de PCR para testar a presença da seqüência nas células contendo mapas de contíguos.

Um tipo de STS se origina a partir de **etiquetas de seqüências expressas** (EST), uma porção de cDNA (DNA complementar, ou seja, uma seqüência de DNA derivada do mRNA de genes expressos). A seqüência contém apenas os éxons do gene, unidos para formar a seqüência que codifica a proteína. Seqüências de cDNA podem ser mapeadas nos cromossomos utilizando a técnica FISH, ou localizadas nos mapas de contíguos.

Como os mapas de contíguos e os sítios com seqüência marcada facilitam a identificação de genes? Se você estiver trabalhando com um organismo que não tenha o seu genoma completo seqüenciado, mas que tenha mapas de contíguos completos disponíveis para todos os cromossomos, você identificaria marcadores STS fortemente ligados ao seu gene, e então localizaria esses marcadores nos mapas de contíguos.

Selecionando genes em genomas

Programas de computadores para a análise de genomas identificam **fases de leitura aberta** ou ORFs. Uma ORF é uma região da seqüência de DNA que começa com um códon iniciador (ATG) e termina com um códon de parada. Uma ORF é uma região codificadora de proteína em potencial.

Técnicas para a identificação de regiões codificadoras de proteínas combinam ou escolhem entre duas abordagens possíveis:

(1) **Detecção de regiões similares a regiões codificadoras conhecidas de outros organismos.** Essas regiões podem codificar seqüências de aminoácidos similares a proteínas já conhecidas, ou ser similares a etiquetas de seqüências

expressas (ESTs). Como as ESTs são derivadas de mRNA, elas correspondem a genes sabidamente transcritos. É necessário seqüenciar apenas algumas centenas de bases iniciais do cDNA para se obter informações suficientes para a identificação de um gene: a caracterização de genes por ESTs é como a indexação de poemas ou de músicas por meio dos seus primeiros versos.

(2) **Métodos *ab initio* que procuram identificar genes a partir de propriedades intrínsecas das seqüências de DNA**. A anotação de genomas assistida por computador é mais completa e acurada para bactérias do que para eucariotos. Os genes de bactérias são relativamente fáceis de identificar porque são contíguos – não apresentam os íntrons característicos dos genomas de eucariotos – e o espaçamento entre os genes é pequeno. Nos organismos superiores, a identificação de genes é mais complexa. Identificar éxons é um dos problemas, uni-los ou recompô-los (*splicing*) é outro. Padrões de recomposição alternativa representam uma dificuldade extra.

Uma estrutura para a identificação de genes em genomas eucarióticos pela metodologia *ab initio* inclui as seguintes características:

- O primeiro éxon (5') começa com um ponto inicial de transcrição, precedido por uma região promotora como a caixa TATA cerca de 30 pb acima (sentido 5') do ponto inicial. Ele não apresenta códons de parada e termina imediatamente antes do sinal de recomposição caracterizado pelo dímero GT. (Ocasionalmente, um éxon não-codificador precede o éxon que contém o códon iniciador.)
- Éxons internos, assim como os iniciais, não apresentam códons de parada. Eles começam imediatamente após o sinal de recomposição caracterizado pelo dímero AG e terminam imediatamente antes do sinal GT.
- O éxon final (3') começa imediatamente após o sinal de recomposição AG e termina com um códon de parada, seguido de uma seqüência sinalizadora poliadenilada. (Ocasionalmente, um éxon não-codificador aparece logo após o éxon que contém o códon de parada.)

Todas as regiões codificadoras possuem características de seqüências não aleatórias baseadas parcialmente em preferências de uso de códons. Empiricamente, sabe-se que estatísticas que empregam hexanucleotídeos têm melhor desempenho em distinguir regiões codificadoras de regiões não-codificadoras. Partindo-se de um conjunto de genes conhecidos de um organismo como conjunto de treinamento, programas de reconhecimento de padrões podem ser calibrados para genomas de interesse.

A detecção precisa de genes é um componente crucial para a análise de seqüências de genomas. Este problema tem destaque importante nas pesquisas atuais.

Genomas de procariotos

A maioria das células procarióticas contém seu material genético na forma de uma única molécula de DNA dupla fita circular grande, geralmente com menos de 5 Mpb. Além dessa molécula principal, essas células também contêm plasmídeos.

As regiões codificadoras de proteínas nos genomas bacterianos não contêm íntrons. Em muitos genomas procarióticos, tais regiões estão parcialmente organizadas em *óperons* – genes adjacentes que são transcritos em uma única molécula de mRNA, sob o mesmo controle transcricional. Em bactérias, os genes de muitos óperons codificam proteínas com funções correlatas. Por exemplo, os genes suces-

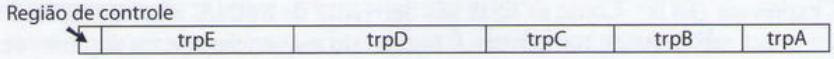


Figura 2.1 O óperon trp de *E. coli* começa com uma região de controle contendo promotor, operador e seqüências líderes. Cinco genes estruturais codificam proteínas que catalisam etapas sucessivas na síntese do aminoácido triptofano a partir do seu precursor, o corismato:

corismato → antranilato → fosforribosil-antranilato → indolglicerol-fosfato → indol → triptofano
(1) (2) (3) (4) (5)

Etapa (1) da reação: trpE e trpD codificam dois componentes da antranilato sintase. Esta enzima tetramérica, formada por duas cópias de cada subunidade, catalisa a conversão de corismato em antranilato. Etapa (2) da reação: a proteína codificada pelo gene trpD também catalisa a fosforribosilação subseqüente do antranilato. Etapa (3) da reação: trpC codifica uma outra enzima bifuncional, fosforribosil-antranilato isomerase – indolglicerol-fosfato sintase. Ela converte o fosforribosil-antranilato em indolglicerol-fosfato através do intermediário carboxifenilamino-deoxirribulose-fosfato. Etapas (4) e (5) da reação: trpB e trpA codificam as subunidades β e α, respectivamente, de uma terceira enzima bifuncional, a triptofano sintase (um tetrâmero $\alpha_2\beta_2$). Um canal na estrutura dessa enzima transporta, sem liberar no solvente, o intermediário da reação produzido pela subunidade α – indolglicerol-fosfato → indol – para o sítio ativo da subunidade β – a qual converte indol em triptofano.

Um gene à parte, trpR, que não está ligado a este óperon, codifica o repressor da via. O repressor pode ligar-se ao DNA na seqüência do operador (localizada na região controladora) apenas quando ligado ao triptofano. A ligação do repressor bloqueia o acesso da RNA polimerase ao promotor, desativando esta via quando o triptofano está presente em abundância. Um controle adicional da transcrição, em resposta aos níveis de triptofano, também é exercido pelo elemento atenuador no mRNA, localizado na seqüência líder. A região atenuadora (1) contém dois códons adjacentes para o aminoácido triptofano e (2) pode adotar estruturas secundárias alternativas, uma das quais bloqueia a transcrição. Os níveis de triptofano regulam os níveis de trp-tRNAs, que, por sua vez, regulam a taxa de progresso dos códons trp adjacentes através do ribossomo. A pausa no ribossomo, na região do mRNA que contém os códons trp adjacentes, em resposta à concentração dos níveis de triptofano, reduz a formação da estrutura secundária do mRNA que termina a transcrição.

sivos do óperon trp de *E. coli* codificam as proteínas que catalisam etapas sucessivas na biossíntese de triptofano (ver Figura 2.1). Em arqueobactérias, uma relação metabólica entre genes em óperons é observada com menor freqüência.

Um genoma procariótico típico contém apenas uma quantidade relativamente pequena de DNA não-codificador (em comparação aos eucariotos), distribuída por toda a seqüência. Em *E. coli* apenas ~11% do seu DNA não é codificador.

O genoma da bactéria *Escherichia coli*

A cepa K-12 de *E. coli* tem sido utilizada por muito tempo como organismo modelo da biologia molecular. O genoma da cepa MG1655, publicado em 1997 pelo grupo de F. Blattner da Universidade de Wisconsin, apresenta 4.639.221 pb em uma única molécula de DNA circular, sem plasmídeos. Aproximadamente 89% da seqüência codifica proteínas ou RNAs estruturais. Um descrição minuciosa revelou:

- 4.284 genes codificadores de proteínas

- 122 genes para RNAs estruturais
- seqüências repetidas não-codificadoras
- elementos reguladores
- sinais de transcrição/tradução
- transposases
- vestígios de prófagos
- seqüências de inserção
- segmentos de composições incomuns, provavelmente elementos exógenos introduzidos por transferência horizontal

A análise da seqüência do genoma exigiu a identificação e a anotação dos genes que codificam proteínas e de outras regiões funcionais. Muitas proteínas de *E. coli* já eram conhecidas antes que o seqüenciamento do seu genoma estivesse completo, fruto de muitos anos de intensos estudos: 1.853 proteínas já haviam sido descritas antes da publicação da seqüência do genoma. Outros genes tiveram sua função atribuída pela identificação de homólogos por meio de buscas em bancos de dados de seqüências. Quanto menor a variação na especificidade da função dos homólogos, mais precisa poderia ser a anotação do genoma. Atualmente, mais de 80% das proteínas podem ter, ao menos, uma função geral atribuída (ver Quadro Distribuição das proteínas de *E. coli* entre 22 grupos funcionais). Outras regiões do genoma são reconhecidas como sítios reguladores ou elementos genéticos móveis, também com base na similaridade com seqüências homólogas conhecidas em outros organismos.

Distribuição das proteínas de *E. coli* entre 22 grupos funcionais

Classe funcional	Número	%
Função reguladora	45	1,05
Proteínas reguladoras putativas	133	3,10
Estrutura celular	182	4,24
Proteínas de membrana putativas	13	0,30
Proteínas estruturais putativas	42	0,98
Fagos, transposons, plasmídeos	87	2,03
Proteínas de ligação e transporte	281	6,55
Proteínas de transporte putativas	146	3,40
Metabolismo energético	243	5,67
Replicação, recombinação, modificação e reparo do DNA	115	2,68
Transcrição, síntese de RNA, metabolismo e modificação	55	1,28
Tradução, modificação pós-traducional de proteínas	182	4,24
Processos celulares (incluindo adaptação e proteção)	188	4,38

→

Distribuição das proteínas de ***E. coli*** entre 22 grupos funcionais (*continuação*)

Biossíntese de co-fatores, grupos prostéticos e carreadores	103	2,40
Chaperonas putativas	9	0,21
Biossíntese e metabolismo de nucleotídeos	58	1,35
Biossíntese e metabolismo de aminoácidos	131	3,06
Metabolismo de ácidos graxos e fosfolipídeos	48	1,12
Catabolismo de compostos de carbono	130	3,03
Metabolismo intermediário central	188	4,38
Enzimas putativas	251	5,85
Outros genes conhecidos (proteína ou fenótipo conhecido)	26	0,61
Hipotético, não classificado, ou desconhecido	1.632	38,06

Adaptado de Blattner *et al.* (1997), The complete genome sequence of *Escherichia coli* K12, *Science*, **277**:1453-1462.

A distribuição dos genes que codificam proteínas no genoma de *E. coli* não parece seguir nenhuma regra simples, seja ao longo da cadeia de DNA ou em fitas diferentes. Na verdade, a comparação de cepas diferentes sugere que os genes são móveis.

O genoma de *E. coli* é relativamente denso em genes. Os genes que codificam proteínas ou RNAs estruturais ocupam ~89% da seqüência. O tamanho médio das ORFs é de 317 aminoácidos. Se os genes estivessem distribuídos de forma homogênea, o tamanho médio das seqüências intergênicas seria de 130 pb; a distância média observada entre os genes é de 118 pb. Contudo, o tamanho destas regiões varia consideravelmente. Algumas são bastante extensas e contêm elementos de função reguladora e seqüências repetidas. A região intergênica mais longa, de 1.730 pb, contém repetições de seqüências não-codificadoras.

Aproximadamente três quartos das unidades transcritas contêm apenas um gene; o restante contém vários genes consecutivos, ou óperons. Estima-se que o genoma de *E. coli* contenha entre 630 e 700 óperons. Os óperons variam em tamanho, embora poucos contenham mais de cinco genes. Os genes em um óperon tendem a apresentar funções relacionadas.

Em alguns casos, a mesma seqüência de DNA codifica partes de mais de uma cadeia polipeptídica. Um gene codifica tanto a subunidade τ quanto a γ da DNA polimerase III. A tradução do gene completo forma a subunidade τ. A subunidade γ corresponde aos dois terços da porção N-terminal da subunidade τ. Uma alteração na fase de leitura neste ponto no ribossomo leva à terminação da cadeia em 50% das vezes, originando uma taxa de expressão de 1:1 das subunidades τ e γ. Não parece haver genes sobrepostos nos quais diferentes fases de leitura codificam proteínas expressas.

Em outros casos, a mesma cadeia polipeptídica aparece em mais de uma enzima. Uma proteína com função própria, a lipoato desidrogenase, é também uma subunidade essencial da piruvato desidrogenase, 2-oxoglutarato desidrogenase e do complexo de clivagem da glicina.

De posse do genoma completo, podemos examinar o repertório de proteínas de *E. coli*. A maior classe de proteínas são as enzimas – aproximadamente 30%

do total de genes. Muitas funções enzimáticas são compartilhadas por mais de uma proteína. Alguns destes grupos de enzimas funcionalmente similares estão proximamente relacionados e parecem ter surgido por duplicação, na própria *E. coli* ou em algum ancestral. Outros grupos de enzimas funcionalmente similares possuem seqüências bastante diferentes e diferem quanto à sua especificidade, regulação ou localização intracelular.

Diversas características desta abundância de enzimas expressas em *E. coli* lhe conferem uma competência metabólica ampla e versátil, permitindo o seu crescimento e competição sob condições variadas:

- Ela pode sintetizar todos os componentes de proteínas e ácidos nucléicos (aminoácidos e nucleotídeos) e co-fatores.
- Ela possui flexibilidade metabólica: pode crescer tanto em meio aeróbio como anaeróbio, utilizando diferentes vias metabólicas para a obtenção de energia. Ela pode crescer em diferentes fontes de carbono. Nem todas as vias metabólicas estão ativas todo o tempo, uma versatilidade que permite resposta a alterações nas condições do ambiente.
- Mesmo para reações metabólicas específicas, existem muitos casos de enzimas múltiplas. Esta redundância contribui para a habilidade de ajustar o metabolismo a condições variadas.
- Entretanto, a *E. coli* não possui um conjunto completo de capacidade enzimática. Ela não pode fixar CO_2 nem N_2.

Descrevemos aqui algumas das características *estáticas* do genoma de *E. coli*, e seu repertório de proteínas. Pesquisas atuais estão enfocando os aspectos dinâmicos, investigando o padrão de expressão de proteínas ao longo do tempo e espaço.

O genoma da arqueobactéria *Methanococcus jannaschii*

S. Luria sugeriu certa vez que, para determinar as características comuns a todas as formas de vida, não se deveria tentar examinar todos os organismos vivos, mas, sim, identificar os organismos mais diferentes de nós mesmos e ver o que temos em comum. A hipótese era que, para se conseguir tal feito, seria necessário encontrar um organismo adaptado a um ambiente o mais diferente possível do nosso.

A exploração do fundo do mar revelou um ambiente tão diferente do que nos é familiar quanto aqueles criados pela ficção científica. Fendas ou fumarolas hidrotérmicas são vulcões subaquáticos que emitem lava e gases aquecidos através de rachaduras no fundo do oceano. Elas criam nichos para comunidades de organismos vivos que não têm contato com a superfície e que dependem dos minerais expelidos pelas fendas hidrotérmicas como sua fonte de nutrientes inorgânicos. Essas fendas mantêm comunidades de microrganismos que são as únicas formas de vida conhecidas que não dependem da luz solar, direta ou indiretamente, como sua fonte de energia.

O microrganismo *Methanococcus jannaschii* foi coletado de uma fenda hidrotérmica a 2.600 m de profundidade na costa da Baja Califórnia, México, em 1983. Ele é um organismo termófilo, que sobrevive em temperaturas entre 48°C e 94°C, com um ótimo a 85°C. É um anaeróbio estrito, capaz de auto-reprodução a partir de componentes inorgânicos. Sua equação metabólica geral é a síntese de metano a partir de H_2 e CO_2.

M. jannaschii pertence ao domínio Archaea, uma das três divisões principais dos seres vivos, juntamente com os domínios Bacteria e Eukarya (ver Figura 1.2). Os Archaea compreendem os grupos dos procariotos, incluindo organismos adaptados a condições ambientais extremas, tais como altas temperatura e pressão, e alta concentração salina.

O genoma de *M. jannaschii* foi seqüenciado em 1996 pelo The Institute for Genomic Research (TIGR). Foi o primeiro genoma de um organismo Archaea seqüenciado. Ele contém um grande cromossomo formado por uma molécula de DNA circular dupla fita de 1.664.976 pb e dois elementos extracromossômicos de 58.407 e 16.550 pb. Existem 1.743 regiões codificadoras preditas, das quais 1.682 estão no cromossomo, e 44 e 12 estão nos elementos extracromossômicos grande e pequeno, respectivamente. Alguns genes de RNA contêm íntrons. Assim como nos demais genomas procarióticos, existe pouco DNA não-codificador.

M. jannaschii parecia satisfazer o objetivo de S. Luria de encontrar nosso parente vivo mais distante. A comparação da seqüência genômica com a de outros organismos mostrou que *M. jannaschii* tem pouca relação com outras formas de vida. Apenas 38% de suas ORFs puderam ter função inferida com base na homologia com proteínas conhecidas de outros organismos. Entretanto, para a grande surpresa de todos, os Archaea são, em alguns aspectos, mais relacionados com os eucariotos do que com as bactérias! Eles são uma mistura complexa. Nos Archaea, as proteínas envolvidas na transcrição, tradução e regulação são mais similares às dos eucariotos. As proteínas envolvidas no metabolismo são mais similares àquelas das bactérias.

O genoma de um dos organismos mais simples: *Mycoplasma genitalium*

O *Mycoplasma genitalium* é uma bactéria infecciosa que causa uretrite não-gonocócica. Seu genoma foi seqüenciado em 1995 por meio de colaboração entre o TIGR, a University of Johns Hopkins e a University of North Carolina. O genoma é composto por uma única molécula de DNA formada por 580.070 pb. Este é o menor genoma celular já seqüenciado. Por enquanto, *M. genitalium* é o que temos de mais próximo a um organismo mínimo, a menor forma de vida independente. (Os vírus, ao contrário, necessitam da maquinaria celular de seus hospedeiros.)

O genoma é denso em regiões codificadoras de proteínas, com 468 genes identificados como proteínas expressas. Algumas regiões do genoma são ricas em genes e outras mais pobres, mas, em geral, 85% do genoma codifica proteínas. O tamanho médio das regiões codificadoras é de 1.040 pb. Assim como em outros genomas bacterianos, as regiões codificadoras não contêm íntrons. Uma compressão ainda maior do genoma é obtida por meio da sobreposição de genes. Parece que muitas dessas sobreposições surgiram devido à perda de códons de parada.

O repertório de genes de *M. genitalium* inclui alguns que codificam proteínas essenciais para a reprodução independente, tais como aqueles envolvidos na replicação, transcrição e tradução do DNA, mais RNAs ribossômicos e transportador. Outros genes são específicos para a atividade infecciosa, incluindo adesinas que medeiam a ligação às células infectadas, moléculas de defesa contra o sistema imune do hospedeiro e proteínas de transporte. Como adaptação ao estilo de vida parasitário do organismo, houve uma grande perda de enzimas metabólicas, incluindo aquelas responsáveis pela biossíntese de aminoácidos – na verdade, um aminoácido está ausente em todas as proteínas do *M. genitalium* (ver *Weblema* 2.7).

Genomas de eucariotos

É raro na ciência encontrar um mundo completamente novo, contendo fenômenos totalmente inesperados. A complexidade do genoma eucariótico é um destes mundos (ver Quadro Inventário de um genoma eucariótico).

Inventário de um genoma eucariótico

DNA moderadamente repetitivo

- Funcional
  - famílias de genes dispersos
    - exemplo: actina, globina
  - arranjos de famílias de genes em *tandem*
    - genes de rRNAs (250 cópias)
    - genes de tRNAs (50 sítios com 10 a 100 cópias cada, nos humanos)
    - genes de histonas em muitas espécies
- Sem função conhecida
  - elementos nucleares intercalados curtos (SINEs)
    - seqüências Alu são um exemplo
    - contêm de 200 a 300 pb
    - centenas de milhares de cópias (300.000 Alu)
    - localizações dispersas (não como repetições em *tandem*)
  - elementos nucleares intercalados longos (LINEs)
    - contêm de 1 a 5 kpb
    - contêm de 10 a 10.000 cópias por genoma
  - pseudogenes

DNA altamente repetitivo

- Minissatélites
  - compostos por repetições de segmentos de 14 a 500 bp
  - contêm de 1 a 5 kpb
  - há muitos tipos diferentes
  - espalhados por todo o genoma
- Microssatélites
  - compostos por repetições de até 13 pb
  - contêm centenas de kpb
  - aproximadamente 10^6 cópias por genoma
  - maior parte da heterocromatina em volta do centrômero
- Telômeros
  - contêm uma unidade de repetição curta (tipicamente de 6 pb: TTAGGG no genoma humano, TTGGGG em *Paramecium*, TAGGG em tripanossomos e TTTAGGG em *Arabidopsis*)
  - contêm de 250 a 1.000 repetições no final de cada cromossomo

Em células eucarióticas, a maior parte do DNA está no núcleo, separado em feixes de nucleoproteínas, os cromossomos. Cada cromossomo contém uma única molécula de DNA fita dupla. Quantidades menores de DNA são observadas em organelas – mitocôndrias e cloroplastos. As organelas surgiram como parasitas intracelulares. Os genomas das organelas geralmente têm a forma de uma cadeia de DNA fita dupla circular, porém, algumas vezes, são lineares e também aparecem como múltiplos círculos. O código genético usado para a tradução dos genes das organelas é diferente do código dos genes nucleares.

Os genomas nucleares de espécies diferentes variam bastante em tamanho (ver páginas 88-89). A correlação entre tamanho de genoma e complexidade do organismo é bastante fraca. Ela certamente não comporta qualquer preconceito de que os homens sejam o ápice da evolução. Em muitos casos, as diferenças nos tamanhos dos genomas refletem as diferenças nas quantidades de seqüências repetitivas.

Além da variação no conteúdo de DNA, as espécies eucarióticas também variam quanto ao número de cromossomos e à distribuição de genes entre eles. Algumas destas diferenças na distribuição de genes envolvem translocações, ou fragmentação ou junção dos cromossomos. Por exemplo, os humanos possuem 23 pares de cromossomos; os chimpanzés têm 24. O cromossomo 2 humano corresponde a uma fusão dos cromossomos 12 e 13 dos chimpanzés (ver Figura 2.2). A dificuldade de pareamento de cromossomos durante a mitose em um zigoto, após um evento de fusão, pode contribuir para o isolamento reprodutivo necessário para a separação de espécies.

Outras diferenças no complemento de cromossomos refletem eventos de duplicação ou hibridização. O trigo utilizado inicialmente na agricultura, no Oriente Médio, pelo menos 10.000 a 15.000 anos atrás, era a variedade diplóide de chamada *einkorn* (*Triticum monococcum*), contendo 14 pares de cromossomos. As variedades *emmer* (*T. dicoccum*), também cultivada desde o paleolítico, e *durum* (*T. turgidum*) são espécies híbridas que se formaram com o cruzamento de parentes da *einkorn* com outras gramíneas selvagens, para produzir espécies tetraplóides. Hibridizações adicionais, com diferentes tipos de trigo selvagem, originaram as formas hexaplóides, incluindo a variedade *spelt* (*T. spelta*) e o trigo comum moderno *T. aestivum*. O *triticale*, um cultivar robusto desenvolvido pela agricultura moderna, e hoje utilizado principalmente na alimentação de animais, é um gênero artificial originado do cruzamento do trigo *durum* (*T. turgidum*) com o centeio (*Secale cereale*). A maioria das variedades de triticale é hexaplóide.

Variedade de trigo	Classificação	Complemento cromossômico
einkorn	*Triticum monococcum*	AA
emmer	*Triticum dicoccum*	AABB
durum	*Triticum turgidum*	AABB
spelt	*Triticum spelta*	AABBDD
comum	*Triticum aestivum*	AABBDD
triticale	*Triticosecale*	AABBRR

A = genoma do trigo diplóide original ou espécie próxima, B = genoma da gramínea selvagem *Aegilops speltoides* ou *Triticum speltoides* ou espécie próxima, D = genoma de outra gramínea selvagem, *Triticum tauschii* ou espécie próxima, R = genoma do centeio *Secale cereale*.

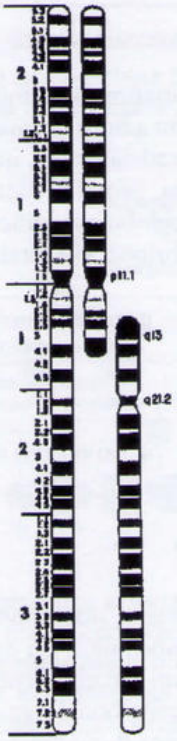


Figura 2.2 Esquerda: cromossomo 2 humano. Direita: cromossomos correspondentes de chimpanzé. (De Yunis, J.J., Sawyer, J.R. & Dunham, K. (1980), The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee, *Science*, **208**:1145-1148.)

Todas essas espécies ainda são cultivadas – algumas em menor quantidade – e possuem suas aplicações específicas na culinária. A espelta, ou *farro* em italiano, é a base de uma sopa bem conhecida; o macarrão é feito do trigo *durum;* e o pão, de *T. aestivum*.

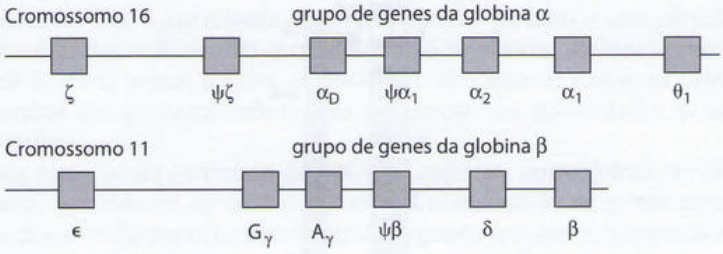
Mesmo em cromossomos individuais, famílias de genes são comuns em eucariotos. Alguns membros da família são parálogos – genes relacionados que divergiram para prover funções diferentes em uma mesma espécie. (Ortólogos, ao contrário, são homólogos que desempenham a mesma função em espécies diferentes. Por exemplo, as globinas α e β humanas são parálogas, e a mioglobina humana e de cavalo são ortólogas.) Outras seqüências relacionadas podem ser pseudogenes, que podem ter surgido por duplicação ou por retrotransposição de um mRNA, seguida pelo acúmulo de mutações a ponto de haver perda da sua função. O grupo de genes da globina em humanos é um bom exemplo (ver Quadro, página 110, O grupo de genes da hemoglobina humana).

O genoma de *Saccharomyces cerevisiae* (fermento de padeiro)

A levedura é um dos organismos eucarióticos mais simples conhecidos. Suas células, como as nossas, contêm um núcleo e outros compartimentos intracelulares especializados. O seqüenciamento do seu genoma, realizado por um consórcio internacional envolvendo cerca de 100 laboratórios, foi completado em 1996.

O grupo de genes da hemoglobina humana

Os genes e pseudogenes humanos da hemoglobina aparecem em grupos nos cromossomos 11 e 16. Um adulto normal sintetiza principalmente três tipos de cadeias de globina: cadeias α e β, que formam o tetrâmero da hemoglobina $\alpha_2\beta_2$; e mioglobina, proteína monomérica encontrada no músculo. Outras formas de hemoglobina, codificadas por genes diferentes, são sintetizadas nos estágios embrionário e fetal.



O grupo de genes da globina α, no cromossomo 16, estende-se por 28 kpb. Ele contém três genes funcionais: ζ e dois genes α idênticos nas suas seqüências codificadoras, e um terceiro gene expresso; dois pseudogenes, ψζ e ψα_1, e ainda um outro gene homólogo de função ainda não definida, θ_1. O grupo de genes da globina β, no cromossomo 11, estende-se por 50 kpb. Ele inclui cinco genes funcionais: ϵ, dois genes γ (G_γ e A_γ), com diferença em apenas um aminoácido, δ e β; e um pseudogene, ψβ. O gene da mioglobina não está ligado a nenhum destes grupos.

Todos os genes humanos da hemoglobina e mioglobina possuem a mesma estrutura de íntron/éxon. Eles contêm três éxons separados por dois íntrons:

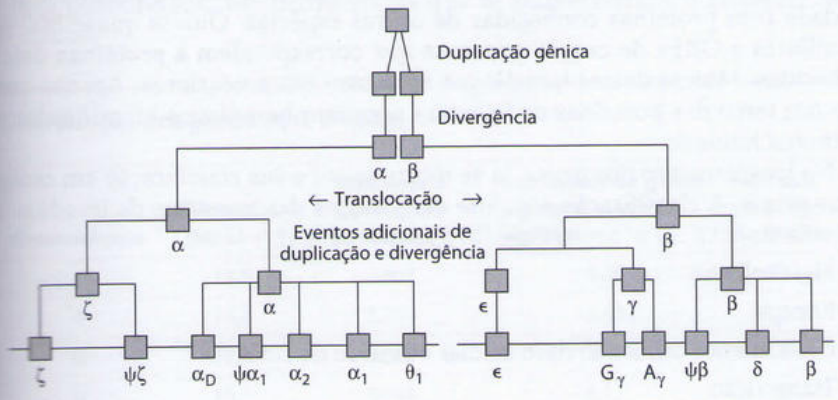
Aqui, E = éxon e I = íntron. Os tamanhos das regiões neste diagrama refletem a estrutura do gene da globina β humana. O padrão de estrutura íntron/éxon é conservado na maioria dos genes de globina expressos nos vertebrados, incluindo as cadeias α e β da hemoglobina e mioglobina. Diferentemente, os genes de globinas de plantas possuem um íntron adicional, genes para globina em *Paramecium* apresentam um íntron a menos e genes das globinas de insetos não possuem íntrons. O gene humano da neuroglobina, um homólogo recentemente descoberto e expresso em baixos níveis no cérebro humano, contém três íntrons, assim como os genes das globinas de plantas.

A distribuição dos genes e pseudogenes da hemoglobina nos cromossomos parece refletir uma evolução por duplicação e divergência (ver diagrama na próxima página). A expressão desses genes segue um padrão rígido de desenvolvimento. No embrião (até a 6ª semana após a concepção), duas cadeias de hemoglobina são sintetizadas principalmente – ζ e ϵ – formando o tetrâmero $\zeta_2\epsilon_2$. Entre seis semanas após a concepção até cerca de oito

→

O grupo de genes da hemoglobina humana (*continuação*)

semanas após o nascimento, a hemoglobina fetal – $\alpha_2\gamma_2$ – é a forma predominante. Ela é substituída pela hemoglobina adulta – $\alpha_2\beta_2$.



As talassemias são doenças genéticas associadas com genes de hemoglobina defectivos ou removidos. A maioria dos indivíduos brancos possui quatro genes para a cadeia α da hemoglobina adulta normal, dois alelos para cada um dos genes α_1 e α_2. Dessa forma, as talassemias α podem ter variações clínicas com diferentes graus de severidade, dependendo de quantos genes expressam as cadeias α normais. Apenas as deleções que resultam em menos de dois genes ativos se apresentam como sintomáticas sob condições normais. Os defeitos genéticos observados incluem a deleção dos dois genes (um processo provavelmente derivado do rearranjo dos genes em *tandem* e das seqüências repetitivas que realizam *crossing-over* com maior probabilidade); e a perda dos sinais de término das cadeias, o que origina um transcrito "contínuo", criando cadeias polipeptídicas mais longas e instáveis.

As talassemias β geralmente são mutações pontuais, incluindo mutações não-sinônimas (substituição de aminoácidos), ou mutações sem sentido (alteração de um tripleto que codifica um aminoácido por um códon de parada), levando a uma terminação prematura ou a uma proteína truncada, mutações nos sítios de recomposição ou mutações nas regiões reguladoras. Algumas deleções incluindo o códon de terminação normal e a região intergênica entre os genes δ e β formam proteínas δ-β fusionadas.

O genoma de levedura contém 12.057.500 pb de DNA nuclear, distribuídos em 16 cromossomos. Os tamanhos dos cromossomos variam em até uma ordem de grandeza, de 1.352 pb do cromossomo IV a 230 kpb do cromossomo I.

O genoma de levedura contém 5.885 genes preditos que codificam proteínas, cerca de 140 genes para RNAs ribossomais, 40 genes para RNAs nucleares pequenos e 275 genes para RNAs transportadores. Em dois aspectos o genoma de levedura é mais denso em regiões codificadoras do que os genomas conhecidos dos eucariotos mais complexos *Caenorhabditis elegans*, *Drosophila melanogaster*, e do homem: (1) Íntrons são relativamente raros e relativamente pequenos. Em levedura, apenas 231 genes apresentam íntrons. (2) Existem menos seqüências repetitivas em comparação com eucariotos mais complexos.

A duplicação do genoma completo de levedura parece ter ocorrido há cerca de 150 milhões de anos. Essa duplicação foi seguida pela translocação de partes do DNA duplicado e pela perda de uma das cópias da maioria (~92%) dos genes.

Dos 5.885 genes codificadores de proteínas potenciais, 3.408 correspondem a proteínas conhecidas. Cerca de 1.000 ou mais apresentam alguma similaridade com proteínas conhecidas de outras espécies. Outras quase 800 são similares a ORFs de outros genomas que correspondem a proteínas desconhecidas. Muitos destes homólogos aparecem em procariotos. Apenas cerca de um terço das proteínas de levedura possuem homólogos identificados no genoma humano.

No levantamento dos genes, já se mostrou útil a sua classificação em categorias gerais. A classificação seguinte das funções das proteínas de levedura foi obtida de `http://mips.gsf.de/projects/funcat/`:

- Metabolismo
- Energia
- Crescimento celular, divisão celular e síntese de DNA
- Transcrição
- Síntese de proteínas
- Direcionamento de proteínas
- Facilitação do transporte
- Transporte celular e mecanismos de transporte
- Biogênese celular
- Comunicação celular/transdução de sinal
- Resgate celular, defesa celular, envelhecimento e morte celular
- Homeostase iônica
- Organização celular
- Elementos de transposição, proteínas de vírus e plasmídeos
- Não classificadas

A levedura é um organismo-modelo para o desenvolvimento de métodos para a determinação de funções de produtos gênicos. A procura por homólogos tem sido exaustiva e contínua. Existem compilações de mutantes nocaute de cada um de seus genes. (Uma seqüência "código de barras" introduzida em cada um dos mutantes facilita a identificação das células capazes de crescer em condições específicas.) Uma fração surpreendentemente grande dessas cepas de um único gene nocaute são viáveis – 80%. A localização celular e os padrões de expressão estão sendo investigados. Diversos tipos de medidas, incluindo as que são baseadas na ativação da transcrição por pares de proteínas capazes de formar dímeros – o método de dois-híbridos – estão produzindo catálogos de interações interprotéicas.

O genoma de *Caenorhabditis elegans*

O verme nematódeo *Caenorhabditis elegans* entrou na pesquisa biológica na década de 1960, com o interesse determinante de Sydney Brenner. Ele reconheceu o potencial de *C. elegans* como um organismo suficientemente complexo para ser interessante, mas simples o bastante para que fosse realizada uma análise completa, em nível celular, do seu desenvolvimento e de seus circuitos neurais.

O genoma de *C. elegans*, concluído em 1998, foi a primeira seqüência completa de DNA de um organismo multicelular. Seu genoma contém aproximadamente 97 Mpb de DNA distribuídos nos pares de cromossomos I, II, III, IV, V e X (ver Quadro). Não existe cromossomo Y. A distinção dos gêneros em *C. elegans* se dá pelo genótipo XX, um hermafrodita que se autofertiliza, e o genótipo XO, um macho.

Distribuição dos genes de *C. elegans*

Cromossomo	Tamanho (Mpb)	Número de genes para proteínas	Densidade de genes para proteínas (kpb/gene)	Número de genes de tRNAs
I	13,9	2.803	5,06	13
II	14,7	3.259	3,65	6
III	12,8	2.508	5,40	9
IV	16,1	3.094	5,17	7
V	20,8	4.082	4,15	5
X	17,2	2.631	6,54	3

O genoma de *C. elegans* é cerca de oito vezes maior do que o de levedura, e tem aproximadamente três vezes mais genes preditos, 19.099. A densidade gênica é relativamente baixa para um eucarioto, ao redor de um gene a cada 5 kpb de DNA. Os éxons cobrem quase 27% do genoma; os genes contêm uma média de cinco íntrons. Aproximadamente 25% dos genes estão dispostos em grupos de genes relacionados.

Muitas das proteínas de *C. elegans* são comuns a outras formas de vida. Outras parecem ser específicas de nematódeos. Quarenta e dois por cento das proteínas têm homólogos em espécies de outros filos, 34% são homólogas a proteínas de outros nematódeos e 24% não possuem homólogos em outras espécies. Muitas das proteínas foram classificadas de acordo com estrutura e função (ver Quadro).

Diversos tipos de genes para RNAs foram identificados. O genoma de *C. elegans* contém 659 genes para tRNAs, quase metade deles (44%) no cromossomo X. RNAs de *spliceossomos* aparecem em cópias dispersas, geralmente idênticas.

***C. elegans*: 20 domínios protéicos mais comuns**

Tipo de domínio	Número
Quimiorreceptor transmembrana (7x)	650
Domínio proteína cínase eucariótica	410
Dois domínios, dedo de zinco tipo C4	240
Colágeno	170
Receptor transmembrana (7x) da família rodopsina	140

→

C. *elegans*: 20 domínios protéicos mais comuns

Tipo de domínio	Número
Dedo de zinco tipo C2H2	130
Lectina tipo C	120
Motivo de reconhecimento de RNA	100
Dedo de zinco tipo C3HC4 (RING finger)	90
Proteína tirosina fosfatase	90
Repetição anquirina	90
Repetição G-beta do domínio WD	90
Domínio homeobox	80
Canal iônico acionado por neurotransmissor	80
Citocromo P450	80
Helicase C-terminal conservada	80
Álcool desidrogenase de cadeia curta	80
UDP-glicuronosil e UDP-glicosil transferases	70
Domínio tipo EGF	70
Superfamília das imunoglobulinas	70

Fonte: Artigo do consórcio do genoma de *C. elegans*, 11 de dezembro de 1998, *Science*.

(Os *spliceossomos* são estruturas subcelulares especializadas que convertem transcritos de pré-mRNA em mRNA maduro pela remoção de íntrons.) Os RNAs ribossômicos aparecem em um arranjo longo em *tandem* na porção final do cromossomo I. Os RNAs 5S aparecem em um arranjo em *tandem* no cromossomo V. Alguns genes de RNAs aparecem nos íntrons de genes que codificam proteínas.

O genoma de *C. elegans* contém muitas seqüências de repetição. Quase 2,6% do genoma consiste em repetições em *tandem*, e 3,6% do genoma contém repetições invertidas que se localizam preferencialmente nos íntrons, e não entre os genes. Repetições da seqüência hexamérica TTAGGC aparecem em diversos locais do genoma. Existem ainda duplicações simples, envolvendo de centenas a dezenas de milhares de kpb.

O genoma de *Drosophila melanogaster*

A mosca-das-frutas, *Drosophila melanogaster*, tem sido alvo de estudos detalhados de genética e de desenvolvimento por um século. A seqüência do seu genoma, produto da colaboração entre a Celera Genomics e o Projeto Genoma Drosophila de Berkeley, foi anunciada em 1999.

Os seus cromossomos são nucleoproteínas complexas. Cerca de um terço do seu genoma está acomodado em heterocromatina, regiões altamente compactadas e enoveladas (e que, portanto, apresentam forte coloração) flanqueando os centrômeros. Os dois terços restantes são eucromatina, uma forma relativamente não-enovelada e menos compacta. A maioria dos genes ativos está na eucromatina. A heterocromatina em *D. melanogaster* contém muitas repetições em *tandem* da seqüência AATAACATAG, e relativamente poucos genes.

O DNA cromossômico total de *D. melanogaster* contém perto de 180 Mpb. A seqüência publicada consiste na porção eucromatina, com aproximadamente 120 Mpb.

O genoma de *D. melanogaster* está distribuído em cinco cromossomos: três autossomos bastante grandes, um cromossomo Y e um quinto cromossomo muito pequeno que contém apenas cerca de 1 Mpb de eucromatina. Foram identificados 13.601 genes, quase o dobro do observado em levedura, porém menos do que em *C. elegans*, o que talvez fosse uma surpresa. A densidade média dos genes na eucromatina é de um gene a cada 9 kpb; muito menor do que a média de um gene por kpb, tipicamente observada em genomas procarióticos.

Apesar de os insetos não estarem proximamente relacionados aos mamíferos na escala evolutiva, o genoma da *Drosophila* é bastante útil no estudo de doenças humanas. Ele contém homólogos para 289 genes humanos implicados em diversas doenças, incluindo câncer e doenças cardiovasculares, neurológicas, endócrinas, renais, metabólicas e hematológicas. Algumas destas proteínas homólogas têm funções diferentes nas moscas e no homem. Outros genes associados a doenças humanas podem ser inseridos e estudados nas moscas. Por exemplo, o gene para a ataxia espinocerebelar do tipo 3 humana, quando expresso em *Drosophila*, produz uma degeneração das células neuronais similar à observada no homem. Já existem modelos da doença de Parkinson e da malária na mosca-das-frutas.

As regiões não-codificadoras do genoma da *Drosophila melanogaster* devem conter regiões que controlam o padrão de desenvolvimento espaço-temporal. A biologia do desenvolvimento da mosca-das-frutas tem sido estudada intensivamente. É um organismo, portanto, em que o estudo da genômica do desenvolvimento se mostrará extremamente informativo.

O genoma de *Arabidopsis thaliana*

Enquanto planta florífera (angiosperma), a *Arabidopsis thaliana* está bastante distante dos demais eucariotos superiores cujas seqüências genômicas estão disponíveis. É necessária uma análise comparativa para identificar características comuns e específicas.

O genoma de *Arabidopsis thaliana* contém aproximadamente 125 Mpb, dos quais 115,4 Mpb foram divulgados em 2000 pelo consórcio The *Arabidopsis* Genome Initiative. Existem cinco pares de cromossomos contendo 25.498 genes preditos. O genoma é relativamente compacto, com um gene a cada 4,6 kpb em média. Este é um valor intermediário entre os procariotos e a *Drosophila*, e é bastante similar ao genoma de *C. elegans*. Os genes de *Arabidopsis* são relativamente pequenos. Os éxons apresentam tipicamente 250 pb, e os íntrons são relativamente pequenos, com tamanho médio de 170 pb. Genes de plantas tipicamente apresentam maior quantidade de conteúdo GC nas regiões codificadoras.

O genoma de *Arabidopsis thaliana*

	Cromossomo					
	1	2	3	4	5	Total
Tamanho (pb)	29.105.111	19.646.945	23.172.617	17.549.867	25.353.409	115.409.949
Número de genes	6.543	4.036	5.220	3.825	5.874	25.498
Densidade (kpb/gene)	4,0	4,9	4,5	4,6	4,4	
Tamanho médio dos genes	2.078	1.949	1.925	2.138	1.974	

Muitas das proteínas de *Arabidopsis* possuem homólogas em animais, mas alguns sistemas, dentre os organismos superiores, são únicos às plantas, como a produção de parede celular e a fotossíntese. Muitas das proteínas compartilhadas com os animais divergiram bastante desde o último ancestral comum. Outra diferença característica entre plantas e animais é que, nas plantas, 25% dos genes nucleares possuem seqüências sinalizadoras que determinam o transporte de seus transcritos para outras organelas – mitocôndrias e cloroplastos – enquanto apenas 5% dos genes dos animais são marcados para o transporte para as mitocôndrias.

Sinais de eventos de duplicação em larga e em pequena escala aparecem no genoma de *Arabidopsis*. Cinqüenta e oito por cento do seu genoma contêm 24 segmentos duplicados, com 100 kpb ou mais de extensão. Os genes nos segmentos duplicados geralmente são homólogos de função ainda *não* identificada. Entretanto, 4.140 genes (17% do total) aparecem em 1.528 arranjos de famílias de genes em *tandem*, contendo mais de 23 membros. O padrão de duplicação sugere que o ancestral de *Arabidopsis thaliana* sofreu uma duplicação de todo o seu genoma para formar espécies tetraplóides. Que este evento ocorreu há bastante tempo – estimado em 112 milhões de anos – é demonstrado pela alta quantidade de perda de uma das cópias dos genes duplicados e pelo elevado grau de divergência daqueles que permaneceram.

As plantas superiores devem integrar os efeitos de três genomas – o nuclear, o de cloroplastos e o mitocondrial. Os genomas das organelas são muito menores:

Distribuição dos genes de *A. thaliana* entre núcleo e organelas

	Nuclear	Cloroplasto	Mitocôndria
Tamanho (kpb)	125.100	154	367
Genes para proteínas	25.498	79	58
Densidade (kpb/gene)	4,5	1,2	6,25

Muitas proteínas sintetizadas por genes nucleares e transportadas até as organelas parecem ter se originado nas organelas e sido transferidas para o núcleo.

O genoma do *Homo sapiens* (o genoma humano)

Aviso

PESSOAS que tentarem encontrar um motivo nesta narrativa serão processadas; pessoas que tentarem encontrar alguma moral serão banidas; pessoas que tentarem encontrar algum enredo serão alvejadas.

Mark Twain, prefácio do livro *As Aventuras de Huckleberry Finn*

Em fevereiro de 2001, o Consórcio Internacional para o Seqüenciamento do Genoma Humano e a Celera Genomics publicaram, separadamente, rascunhos do genoma humano. A seqüência continha aproximadamente $3{,}2 \times 10^9$ pb, trinta vezes maior do que os genomas de *C. elegans* ou *Drosophila melanogaster*. Uma razão para esta disparidade no tamanho é que as seqüências codificadoras correspondem a menos do que 5% do genoma humano; seqüências de repetições correspondem a mais do que 50% do genoma. Talvez a maior surpresa tenha sido o pequeno número de genes identificados. A descoberta de apenas 20.000

a 25.000 genes sugere que padrões de recomposição alternativos dêem uma contribuição significativa ao nosso repertório de proteínas. Estima-se que aproximadamente 50% dos genes apresentem padrões de recomposição alternativos (*alternative splicing*).

O genoma humano publicado está distribuído em 22 pares de cromossomos, mais os cromossomos X e Y. O conteúdo de DNA dos autossomos varia de 279 Mpb a 48 Mpb. O cromossomo X contém 163 Mpb, e o cromossomo Y, apenas 51 Mpb.

Os éxons dos genes humanos codificadores de proteínas são relativamente pequenos se comparados aos de outros genomas eucarióticos conhecidos. Os íntrons são relativamente maiores. Como conseqüência, os genes que codificam proteínas se estendem por segmentos longos de DNA. Por exemplo, o gene da distrofina, codificando uma proteína de 3.685 aminoácidos, possui um tamanho de mais de 2,4 Mpb.

Genes que codificam proteínas

A análise do repertório de proteínas humanas derivadas da seqüência do genoma tem se mostrado bastante complicada pelos problemas de confiabilidade na detecção de genes e por causa dos padrões de recomposição alternativos.

As principais categorias de proteínas em uma classificação funcional são:

Função	Número	% do genoma
Ligação a ácidos nucléicos	2.207	14,0
Ligação ao DNA	1.656	10,5
Proteína de reparo do DNA	45	0,2
Fator de replicação do DNA	7	0,0
Fator de transcrição	986	6,2
Ligação ao RNA	380	2,4
Proteína estrutural do ribossomo	137	0,8
Fator de tradução	44	0,2
Ligação aos fatores de transcrição	6	0,0
Regulação do ciclo celular	75	0,4
Chaperona	154	0,9
Motora	85	0,5
Ligação à actina	129	0,8
Proteína de defesa/imunidade	603	3,8
Enzima	3.242	20,6
Peptidase	457	2,9
Endopeptidase	403	2,5
Proteína cínase	839	5,3
Proteína fosfatase	295	1,8
Ativador enzimático	3	0,0
Inibidor enzimático	132	0,8
Inibidor de apoptose	28	0,1

Função	Número	% do genoma
Transdução de sinal	1.790	11,4
Receptor	1.318	8,4
Receptor transmembrana	1.202	7,6
Receptor acoplado à proteína G	489	3,1
Receptor olfativo	71	0,4
Proteína de armazenamento	7	0,0
Adesão celular	189	1,2
Proteína estrutural	714	4,5
Proteína estrutural de citoesqueleto	145	0,9
Transporte	682	4,3
Canal iônico	269	1,7
Transportador de neurotransmissor	19	0,1
Proteína de ligação a ligantes ou carreadora	1.536	9,7
Transferência de elétrons	33	0,2
Citocromo P450	50	0,3
Supressor de tumor	5	0,0
Não classificadas	4.813	30,6
Total	15.683	100,0

De: `http://www.ebi.ac.uk/proteome/`
[em Functional classification of *H. sapiens* using Gene Ontology (GO): General Statistics (InterPro proteins with GO hits)]

Uma classificação baseada na estrutura revelou os tipos mais comuns:

Tipos mais comuns de proteínas

Proteína	Número
Imunoglobulina e domínio do complexo de histocompatibilidade principal	591
Dedo de zinco tipo C2H2	499
Proteína cínase eucariótica	459
Superfamília GPCR tipo rodopsina	346
Sítio ativo da família de proteínas serina/treonina cínase	285
Domínio tipo EGF	259
Região de ligação ao RNA RNP-1 (motivo de reconhecimento de RNA)	214
Repetições beta 40 WD da proteína G	196
Domínio 3 de homologia a Src (SH3)	194
Domínio de homologia à pleckstrin (PH)	188
Família EF-hand	185
Domínio homeobox	179

Proteína	Número
Domínio catalítico tirosina cínase	173
Imunoglobulina tipo V	163
RING finger	159
Extensina rica em prolina	156
Domínio fibronectina tipo 3	151
Repetição anquirina	135
Caixa KRAB	133
Subtipo de imunoglobulinas	128
Domínio caderina	118
Domínio PZD (também conhecido como DHR ou GLGF)	117
Repetições ricas em leucina	113
Serina proteases, família tripsina	108
Superfamília GTPase Ras	103
Domínio 2 de homologia a Src (SH2)	100
Domínio BTB/POZ	99
Repetições TPR	92
Superfamília ATPase AAA	92
Sítio de hidroxilação de ácido aspártico e asparagina	91

De: http://ebi.ac.uk/proteome/

Seqüências de repetições

Seqüências de repetições compreendem mais de 50% do genoma:

- Elementos de transposição, ou repetições intercaladas – quase metade de todo o genoma! Elas incluem os LINES e SINES (ver Quadro).

Tipos de elementos de transposição no genoma humano

Elemento	Tamanho (pb)	Número de cópias	Fração do genoma (%)
SINES	100-300	1.500.000	13
LINES	6.000-8.000	850.000	21
Repetições terminais longas	15.000-110.0000	450.000	8
Fósseis de transposons de DNA	80-3.000	300.000	3

- Pseudogenes retropostos.
- Simples "gaguejos" – repetições de pequenos oligômeros. Incluem os minissatélites e os microssatélites. Repetições de trinucleotídeos, tal como CAG, correspondendo a repetições de glutamina na proteína correspondente, estão implicadas em várias doenças.

- Duplicação de segmentos de blocos de ~10 a ~300 kpb. Duplicações intercromossômicas aparecem em cromossomos não-homólogos, às vezes em múltiplos locais. Algumas duplicações intracromossômicas incluem regiões duplicadas próximas e de muitos kpb e de seqüência bastante similar implicadas em doenças genéticas; por exemplo, a síndrome de Charcot-Marie-Tooth do tipo 1A, uma neuropatia periférica progressiva que resulta da duplicação de uma região contendo o gene da proteína 22 da mielina periférica.
- Blocos de repetições em *tandem*, incluindo famílias de genes.

RNA

Genes de RNA no genoma humano incluem:

1. 497 genes de RNA transportador. Um grande grupo contém 140 genes de tRNAs em uma região de 4 Mpb no cromossomo 6.
2. Genes para RNAs ribossomais 28S e 5,8S aparecem em unidades de repetições em *tandem* de 44 kpb, com um número de cópias variando entre 150 e 200. Genes de RNA 5S também aparecem agrupados em *tandem* contendo entre 200 e 300 genes, sendo que o maior destes agrupamentos se encontra no cromossomo 1.
3. RNAs nucleolares pequenos incluem duas famílias de moléculas que clivam e processam os RNAs ribossômicos.
4. snRNAs de *spliceossomos*, incluindo os snRNAs U1, U2, U4, U5 e U6, muitos dos quais aparecem em grupos de repetições em *tandem* de seqüências praticamente idênticas, ou repetições invertidas.

Recursos na *web*: Informações do genoma humano

Acesso interativo a seqüências de DNA e de proteínas:

`http://www.ensembl.org/index.html`
`http://www.genome.ucsc.edu/`

Imagens de cromossomos, mapas e *loci*:

`http://www.ncbi.nlm.nih.gov/genome/guide/human/`

GeneMap'99:

`http://www.ncbi.nlm.nih.gov/projects/genome/genemap99/`

Uma visão geral da estrutura do genoma humano:

`http://hgrep.ims.u-tokyo.ac.jp/`

Polimorfismos de um único nucleotídeo:

`http://snp.cshl.org/`

Doenças genéticas humanas:

`http://www.ncbi.nlm.nih.gov/OMIM`
`http://www.geneclinics.org/profiles/all.html`

Aspectos sociais, éticos e legais:

`http://www.genome.gov/ELSI/`

Polimorfismos de um único nucleotídeo (SNPs)

Um polimorfismo de um único nucleotídeo ou SNP (lê-se "snip") é uma variação genética entre indivíduos, limitada a um único par de bases, o qual pode ser substituído, inserido ou removido. A anemia falciforme é um exemplo de doença causada por um SNP: uma mutação A→T no gene da globina β causa uma mudança Glu→Val, criando uma superfície adesiva na molécula de hemoglobina, o que leva à polimerização da forma deoxi.

Os SNPs estão distribuídos ao longo de todo o genoma, ocorrendo em média a cada 5.000 pares de base. Apesar de terem surgido por mutações, muitas das posições que contêm SNPs apresentam baixas taxas de mutação, e podem ser utilizadas como marcadores moleculares estáveis no mapeamento de genes. Um consórcio formado pelos quatro principais centros de genomas acadêmicos e por onze companhias privadas está elaborando o mapa de SNPs humano, com alta densidade de dados e qualidade, com o comprometimento de que ele seja de acesso público. Uma versão publicada recentemente contém 1,8 milhões de SNPs.

Nem todos os SNPs estão ligados a doenças. Muitos se encontram em regiões não-funcionais do genoma (embora a densidade de SNPs seja maior do que a média nas regiões contendo genes). Alguns SNPs observados nos éxons são mutações para códons sinônimos, ou causam substituições que não afetam de maneira significativa a função da proteína. Outros tipos de SNPs podem causar mais do que uma perturbação local em uma proteína: (1) uma mutação que altere um códon normal para um códon de parada, ou vice-versa, pode causar um truncamento prematuro da síntese da proteína ou, por causa da ausência de um códon de parada, levar à síntese de uma proteína mais longa do que a original. (2) Uma deleção ou inserção acarretará uma mudança de fase na tradução.

Os alelos A, B e O dos genes para os tipos sangüíneos ilustram estas possibilidades. Os alelos A e B diferem entre si pela substituição de quatro SNPs. Eles codificam enzimas relacionadas que adicionam diferentes unidades de sacarídeos a um antígeno na superfície das células vermelhas do sangue.

Alelo	Seqüência	Sacarídeo
A	...gctggtgacccctt...	N-acetilgalactosamina
B	...gctcgtcaccgcta...	galactose
O	...cgtggt-acccctt...	—

O alelo O sofreu uma mutação que causou uma mudança na fase de leitura e não produz enzimas ativas. As células vermelhas do sangue de indivíduos com tipo sangüíneo O não contêm o antígeno A nem o B. É por essa razão que estas pessoas são doadoras universais nas transfusões sangüíneas. A perda da atividade enzimática não parece ter conseqüências adversas. Na verdade, pessoas com tipo sangüíneo B e O apresentam uma resistência maior à varíola.

A correlação entre uma doença e um SNP específico é uma vantagem para a prática clínica, pois torna relativamente fácil a identificação de pessoas afetadas ou portadoras. Porém, se a doença for causada pela disfunção de uma proteína específica, deve haver muitos sítios de mutação responsáveis por sua inativação. Um sítio particular de mutação pode predominar se (1) todos os portadores do gene descendem de um único indivíduo no qual ocorreu a mutação, e/ou (2) a doença for o resultado de um *ganho* e não de uma perda de uma

propriedade específica, tal como a habilidade das hemoglobinas falciformes de se polimerizarem, e/ou (3) a taxa de mutação em um sítio específico for muito alta, como a mutação Gli380→Arg no gene receptor 3 do fator de crescimento de fibroblastos (FGFR3), associada com a acondroplasia (uma síndrome que inclui a baixa estatura).

Ao contrário, muitas mutações independentes foram detectadas nos genes BRCA1 e BRCA2, *loci* associados com uma predisposição aumentada para manifestação precoce dos cânceres de ovário e de mama (ver Capítulo 6). Os produtos normais desses genes funcionam como supressores de tumor. Mutantes que apresentam inserções ou deleções geralmente causam uma alteração na fase de leitura que produz uma proteína inativa ou resulta na perda da expressão dessa proteína. Entretanto, não é possível deduzir de antemão se uma mutação nova, com **substituição** de bases nos genes BRCA1 ou BRCA2, confere maior ou menor predisposição ao câncer.

Tratamentos de doenças causadas por proteínas defeituosas ou ausentes incluem:

1. **Provisão da proteína funcional** Já mencionamos os exemplos da insulina para o diabetes e do fator VIII para a forma mais comum de hemofilia. Outro exemplo é a administração de hormônio do crescimento humano em pacientes com redução severa dos níveis normais ou que não o sintetizam. O uso de proteínas recombinantes elimina os riscos de transmissão de AIDS, através de transfusões de sangue, ou da doença priônica Creutzfeld-Jakob obtida do hormônio de crescimento isolado de extratos brutos de pituitária.
2. **Ajustes ao estilo de vida que tornem a proteína desnecessária** A fenilcetonúria (PKU) é uma doença genética causada pela deficiência da fenilalanina hidroxilase, a enzima que converte fenilalanina em tirosina. O acúmulo de altos níveis de fenilalanina causa defeitos no desenvolvimento, incluindo retardo mental. Os sintomas podem ser evitados com uma dieta livre de fenilalanina. A triagem de recém-nascidos para níveis elevados de fenilalanina no sangue é exigida por lei nos Estados Unidos e em muitos outros países.
3. A terapia gênica para a substituição da proteína ausente ou mutante é um campo ativo de pesquisa.

Outras aplicações clínicas dos SNPs refletem correlações entre genótipo e reação à terapia (farmacogenômica). Por exemplo, um SNP no gene da N-acetil transferase (NAT-2) está relacionado com a neuropatia periférica – fraqueza, dormência e dores nos braços, pernas, mãos ou pés – como efeito colateral do tratamento com isoniazida (hidrazida do ácido isonicotínico), um tratamento comum para a tuberculose (ver Capítulo 6). Pacientes cujos testes são positivos para este SNP recebem tratamentos alternativos.

Diversidade genética na antropologia

As informações obtidas com os SNPs são de grande utilidade na antropologia, fornecendo dados sobre variações históricas no tamanho da população e padrões de migração.

Os graus de diversidade genética são interpretáveis em termos do tamanho da população fundadora. Fundadores são o conjunto original de indivíduos dos quais uma população inteira descende. Esses fundadores podem ser os colonizadores originais, como os polinésios que primeiro se estabeleceram na Nova Zelândia, ou simplesmente os sobreviventes de um evento de quase extinção.

Os chitas, ou leopardos de caça da Índia, apresentam os efeitos de um gargalo em uma população, que se estima tenha ocorrido a 10.000 anos atrás. Todos os chitas vivos estão tão fortemente relacionados uns aos outros como o são irmãos em outras espécies. Extrapolações da variação do DNA mitocondrial no homem contemporâneo sugerem um único ancestral materno, o qual viveu entre 14.000 e 200.000 anos atrás. Chamar esse ancestral comum de Eva sugere que ela tenha sido a primeira mulher. Porém, essa Eva Mitocondrial era, de fato, a fundadora de uma população que sobreviveu à extinção de descendentes de todos os contemporâneos da sua espécie.

Os SNPs específicos de uma população podem fornecer informações sobre migrações. Seqüências mitocondriais fornecem informações sobre ancestrais femininos e seqüências do cromossomo Y fornecem informações sobre os ancestrais masculinos. Por exemplo, foi sugerido que a população da Islândia – habitada inicialmente a cerca de 1.100 anos atrás – descende de homens oriundos da Escandinávia e de mulheres oriundas tanto da Escandinávia quanto das Ilhas Britânicas. Escritos medievais islandeses se referem a invasões a povoados nas Ilhas Britânicas.

Uma relação fascinante entre seqüências de DNA humano e famílias de línguas foi investigada por L. L. Cavalli-Sforza e colegas. Estes estudos se mostraram úteis na determinação das inter-relações entre as línguas faladas pelos índios norte-americanos. Este grupo também confirmou que os bascos, uma população que apresenta um já reconhecido isolamento lingüístico, também são geneticamente isolados.

Na descoberta de populações isoladas, a genética antropológica fornece informações úteis na medicina, pois o mapeamento de genes relacionados com doenças se torna fácil se as variações espontâneas forem baixas. As populações européias geneticamente isoladas são, além dos bascos, os finlandeses, os islandeses, os galeses e os lapões.

Diversidade genética e identificação pessoal

Variações em nossa seqüência de DNA nos conferem marcas individuais, como impressões digitais, úteis na identificação e no estabelecimento de relações, incluindo mas não limitadas às questões de paternidade. O uso da análise de DNA como evidência em processos criminais está bem estabelecido agora.

Técnicas de impressão "digital" genética eram originalmente baseadas nos padrões das repetições em *tandem* de número variável (VNTRs), mas foram expandidas para incluir a análise de outras características, como as seqüências de DNA mitocondrial.

Para a maioria de nós, todas as nossas mitocôndrias são geneticamente idênticas, uma condição chamada de homoplasmia. Contudo, alguns indivíduos contêm mitocôndrias com seqüências diferentes de DNA; isso é heteroplasmia. Esta variação de seqüências, quando em um gene associado a uma doença, pode complicar a identificação dos padrões de herança da doença.

O caso mais famoso de heteroplasmia envolveu o czar Nicholas II da Rússia. Após a revolução de 1917, o czar e sua família foram exilados em Yekaterinburg, na Rússia central. Na noite de 16 para 17 de julho de 1918, o czar, a czarina Alexandra e pelo menos três de seus cinco filhos, mais seu médico e três empregados que haviam acompanhado a família no exílio, foram assassinados, e seus corpos enterrados em uma cova secreta. Quando os restos dos corpos foram encontrados, a montagem dos ossos e os exames das arcadas dentárias sugeriram – e a análise das seqüências de DNA confirmaram – que se tratava de

um grupo de uma mesma família. A identificação dos restos mortais da czarina foi comprovada pela comparação da sua seqüência de DNA mitocondrial com a de um parente materno, o príncipe Philip, chanceler da Universidade de Cambridge, duque de Edimburgo e sobrinho-neto da czarina.

No entanto, a comparação das seqüências de DNA mitocondrial dos restos mortais do czar Nicholas II com as seqüências de dois parentes maternos revelou uma diferença na base 16.169: o czar apresentava uma citosina (C) e seus parentes uma timina (T). Questões políticas e mesmo religiosas de extrema sensibilidade não permitiam que dúvidas pudessem existir. Testes adicionais mostraram que o czar era heteroplásmico; a base T era um componente menor das suas seqüências de DNA mitocondrial na posição 16.169. Para confirmar a identidade e eliminar qualquer dúvida, o corpo do Grão-duque Georgij, irmão do czar, foi exumado e também apresentou essa mesma heteroplasmia.

Análise genética da domesticação do gado

Os recursos animais são um aspecto integral e essencial da cultura humana. Análises das seqüências de DNA iluminaram alguns aspectos do seu desenvolvimento histórico e das variedades genéticas que caracterizam as populações modernas.

O gado doméstico contemporâneo inclui as espécies familiares no oeste da Europa e norte da América, *Bos taurus*; e o zebu encontrado na África e Índia, *Bos indicus*. A diferença mais óbvia nas características físicas é a corcunda do zebu. Acredita-se que a domesticação do gado ocorreu uma única vez, entre 8.000 e 10.000 anos atrás, e que as duas espécies divergiram após a domesticação.

Análises das seqüências de DNA mitocondrial do gado europeu, africano e asiático sugerem, porém, que (1) todas as raças européias e africanas estão mais relacionadas entre si do que cada uma delas com as raças indianas, e (2) que os dois grupos divergiram há cerca de 200.000 anos, implicando domesticações recentes independentes de espécies diferentes. As similaridades físicas entre o zebu africano e indiano (e outras similaridades no nível molecular; por exemplo, os marcadores VNTR do DNA nuclear) devem, portanto, ser decorrência da importação de gado da Índia para a África Oriental.

Evolução de genomas

A disponibilidade de informações completas sobre seqüências de genomas alterou o rumo das pesquisas. Um desafio comum na análise de genomas é identificar "eventos interessantes". Uma taxa de mutação espontânea nas seqüências codificadoras se reflete nas substituições **sinônimas** de nucleotídeos: alterações nos códons que não alteram o aminoácido. Utilizando isso como referência, pode-se procurar por casos nos quais existem taxas significativamente mais altas de substituições **não-sinônimas** de nucleotídeos: alterações no códon que causam mutações na proteína correspondente. (Note, todavia, que mudanças sinônimas não são, necessariamente, seletivamente neutras.)

Dado o alinhamento das seqüências de dois genes, podemos calcular K_s, o número de substituições sinônimas, e K_a, o número de substituições não-sinô-

nimas. (O cálculo envolve mais do que uma simples contagem por causa da necessidade de se estimar e corrigir possíveis alterações múltiplas.) Uma razão K_a/K_s alta identifica pares de seqüência claramente apresentando seleção positiva e, também, possivelmente, até mesmo mudanças funcionais.

Esta nova área da genômica comparativa lida com questões que só agora podem ser abordadas, tais como;

- Quais *genes* são compartilhados por filos diferentes? Quais genes são exclusivos a um filo? O arranjo desses genes no genoma varia de um filo para outro?
- Quais *proteínas* homólogas são compartilhadas por filos diferentes? Quais proteínas são exclusivas a um filo? A integração das atividades dessas proteínas varia de um filo para outro? O mecanismo de controle do padrão de expressão dessas proteínas varia de um filo para outro?
- Quais *funções bioquímicas* são compartilhadas por diferentes filos? Quais funções bioquímicas são exclusivas a um filo? A integração dessas funções bioquímicas varia de um filo para outro? Se dois filos compartilham uma função, e a proteína que executa essa função em um filo possui um homólogo no outro filo, as proteínas homólogas executam a mesma função?

As mesmas perguntas poderiam ser aplicadas a diferentes espécies de cada filo.

M. A. Andrade, C. Ouzounis, C. Sander, J. Tamames e A. Valencia compararam o repertório protéico de espécies dos três principais domínios de seres vivos: *Haemophilus influenzae* representando as bactérias, *Methanococcus jannaschii* representando as arqueobactérias e *Saccharomyces cerevisiae* (levedura) representando os eucariotos. As suas classificações das funções das proteínas continham, como categorias principais, processos que envolvem energia, informação, comunicação e regulação:

Classes funcionais gerais

- Energia
  - Biossíntese de co-fatores, aminoácidos
  - Metabolismo central e intermediário
  - Metabolismo energético
  - Ácidos graxos e fosfolipídeos
  - Biossíntese de nucleotídeos
  - Transporte
- Informação
  - Replicação
  - Transcrição
  - Tradução
- Comunicação e regulação
  - Funções reguladoras
  - Envelope celular/parede celular
  - Processos celulares

O número de genes nas três espécies é:

Espécies	Número de genes
Haemophilus influenzae	1.680
Methanococcus jannaschii	1.735
Saccharomyces cerevisiae	6.278

Existem proteínas comuns com funções comuns entre estas espécies? Na categoria energia, os três domínios compartilham proteínas. Na categoria comunicação e regulação, as proteínas são exclusivas a cada domínio. Na categoria informação, o domínio Archaea compartilha algumas proteínas com o domínio Bacteria e outras com o domínio Eukarya.

Análises das funções compartilhadas pelos três domínios levaram pessoas a perguntar se seria possível definir um **organismo mínimo** – isto é, um organismo com o menor genoma possível que fosse consistente com uma forma de vida independente e baseada no dogma central DNA→RNA→proteína (ou seja, excluindo-se as formas de vida livres de proteínas e baseadas apenas em RNA). O organismo mínimo deve ter a capacidade de se reproduzir, mas não é necessário que seja capaz de competir em taxas de crescimento e de reprodução com outros organismos. Pode-se também assumir que esse organismo cresça em um meio rico em nutrientes, o que o exime de responsabilidades de biossíntese e dispensa as funções de resposta ao estresse, como o reparo do DNA.

O organismo conhecido com o menor genoma e vida independente é o *Mycoplasma genitalium*, com 468 seqüências de proteínas preditas. Em 1996, A. R. Mushegian e E. V. Koonin comparam os genomas de *M. genitalium* e de *Haemophilus influenzae*. (Naquela época estavam disponíveis apenas seqüências completas de genomas de bactérias.) O último ancestral comum dessas bactérias bastante divergidas viveu a cerca de dois bilhões de anos atrás. Dos 1.703 genes que codificam proteínas em *H. influenzae*, 240 têm homólogos em *M. genitalium*. Mushegian e Koonin propuseram que estas proteínas devem ser essenciais, mas não necessariamente suficientes para uma forma de vida autônoma – pois algumas funções essenciais devem ser desempenhadas por proteínas não relacionadas nos dois organismos. Por exemplo, o grupo comum de 240 proteínas deixou algumas brechas em vias essenciais, que poderiam ser preenchidas adicionando-se 22 enzimas de *M. genitalium*. Finalmente, a eliminação de redundâncias funcionais e genes específicos de cada parasita resultou em uma lista com 256 genes, propostos como sendo o conjunto mínimo suficiente *e* necessário de genes.

Como é o genoma mínimo proposto? As classes funcionais incluem:

- Tradução, incluindo síntese de proteínas
- Replicação do DNA
- Recombinação e reparo – uma segunda função das proteínas envolvidas na replicação do DNA
- Aparato de transcrição
- Proteínas tipo chaperonas
- Metabolismo intermediário – a via glicolítica
- Ausência da biossíntese de nucleotídeos, aminoácidos ou ácidos graxos
- Maquinaria de exportação de proteínas
- Repertório limitado de proteínas de transporte de metabólitos

Deve ser enfatizado que a viabilidade de um organismo com estas proteínas não foi demonstrada. Além disso, mesmo se experimentos provassem que algum conteúdo genético mínimo – este conjunto proposto ou algum outro – é necessário e suficiente, isso não responde à questão relacionada de identificar o complemento gênico do ancestral comum de *M. genitalium* e *H. influenzae*, ou das formas de vida celular mais primitivas. Não obstante, apenas 71% do conjunto mínimo proposto de 256 proteínas têm homólogas reconhecidas entre as proteínas de organismos eucarióticos ou arqueobactérias.

Contudo, a identificação das funções necessariamente comuns a todas as formas de vida nos permite investigar até que ponto diferentes formas de vida desempenham essas funções da mesma maneira. Reações similares são catalisadas por proteínas homólogas em espécies diferentes? Análises de genomas revelaram famílias de proteínas homólogas nos domínios Archaea, Bacteria e Eukarya. A hipótese é de que essas proteínas tenham evoluído a partir de um gene ancestral comum por meio de uma série de eventos de especiação e duplicação, embora alguns casos devam ser efeitos de transferência horizontal. O desafio aqui é mapear funções comuns e proteínas comuns.

Vários milhares de famílias de proteínas homólogas foram identificadas em Archaea, Bacteria e Eukarya. Espécies distintas contêm quantidades diferentes dessas famílias comuns de proteínas: em bactérias, essa quantidade varia desde *Aquifex aeolicus*, no qual 83% das proteínas possuem homólogas em Archaea e eucariotos, até *Borrelia burgdorferi*, com apenas 52% das proteínas com homólogos em Archaea e eucariotos. Genomas do domínio Archaea apresentam, de alguma forma, maior porcentagem (62 a 71%) de proteínas com homólogos em bactérias e eucariotos. Porém, apenas 35% das proteínas de levedura têm homólogos em bactérias e Archaea.

Um conjunto de proteínas comuns executa um conjunto de funções comuns? Entre as proteínas do conjunto mínimo que foram identificadas em *M. genitalium*, apenas cerca de 30% têm homólogas em todos os genomas conhecidas. Outras funções essenciais devem ser desempenhadas por proteínas não relacionadas ou, em alguns casos, por homólogas ainda não identificadas. As famílias de proteínas para as quais homólogas desempenham funções comuns em Archaea, Bacteria e Eukarya são compostas principalmente de proteínas envolvidas na tradução e na biossíntese:

Classe funcional da proteína	Número de famílias identificadas em todos os genomas conhecidos
Tradução, incluindo a estrutura do ribossomo	53
Transcrição	4
Replicação, recombinação, reparo	5
Metabolismo	9
Processos celulares (chaperonas, secreção, divisão celular, biossíntese da parede celular)	9

O quadro que se desenha é que a evolução explorou, em proporções variadas, o vasto potencial das proteínas para diferentes tipos de funções. Esta variação foi mais conservadora na área de síntese de proteínas.

Recursos na *web*: Bancos de dados de famílias de genes alinhados

Pfam: Banco de dados de famílias de proteínas

`http://www.sanger.ac.uk/Software/Pfam/`

COG: Agrupamentos de grupos ortólogos

`http://www.ncbi.nlm.nih.gov/COG/`

HOBACGEN: Banco de dados de genes homólogos de bactérias

`http://pbil.univ-lyon1.fr/databases/hobacgen.html`

HOVERGEN: Banco de dados de genes homólogos de vertebrados

`http://pbil.univ-lyon1.fr/databases/hovergen.html`

TAED: Banco de dados de Evolução Adaptativa

`http://www.bioinfo.no/tools/TAED`

Muitos outros sítios contêm dados sobre famílias individuais.

Por gentileza, passem os genes: transmissão horizontal de genes

Pela observação de que a tripsina de *S. griseus* é mais similar à tripsina bovina do que às proteinases de outros micróbios, Brian Hartley comentou em 1970 que "... a bactéria deve ter sido infectada por uma vaca." Este era um exemplo claro de transferência lateral ou horizontal de genes – uma bactéria assimilando um gene do solo no qual ela estava crescendo, gene que um organismo de outra espécie tinha depositado ali. Outro exemplo são os experimentos clássicos de transformação de pneumococos, realizados por O. Avery, C. MacLeod e M. McCarthy, que identificaram o DNA como sendo o material genético. Em geral, a transferência horizontal de genes é a aquisição de material genético de um organismo por outro organismo, por meios naturais e não por procedimentos laboratoriais, através de mecanismos que não incluam a descendência direta durante a replicação ou reprodução. Diversos mecanismos de transferência horizontal são conhecidos, incluindo a assimilação direta, como nos experimentos de transformação de pneumococos, ou mediante um carreador viral.

A análise das seqüências de genomas mostrou que a transferência horizontal de genes não é um evento raro, mas que afetou a maioria dos genes nos microrganismos. Este fenômeno exige uma mudança em nossa forma de pensar desde a clonagem comum até os modelos de hereditariedade parentais. Evidências de transferência horizontal incluem (1) discrepâncias entre árvores evolucionárias construídas a partir de genes diferentes e (2) comparações diretas entre genes de espécies distintas:

- Em *E. coli*, cerca de 25% dos genes parecem ter sido adquiridos por transferência de outras espécies.
- Na evolução microbiana, a transferência horizontal de genes é mais prevalente entre genes operacionais – aqueles responsáveis pelas atividades de manutenção, como a biossíntese – do que entre os genes informacionais – aqueles responsáveis pelas atividades organizacionais, como transcrição e tradução. Por exemplo: *Bradyrhizobium japonicum*, uma bactéria fixadora de nitrogênio e simbiótica com plantas superiores, possui dois genes para a glutamina sintetase. Um desses genes é similar aos mesmos de seus parentes bacterianos;

o outro tem 50% de identidade com o gene de plantas superiores. Rubisco (ribulose-1,5-bifosfato carboxilase-oxigenase), a enzima que primeiro fixa o dióxido de carbono no início do ciclo de Calvin da fotossíntese, foi transferida entre bactérias, mitocôndrias e plastídeos de algas, além de sofrer duplicação gênica. Muitos genes de fagos encontrados no genoma de *E. coli* fornecem exemplos adicionais e sugerem um mecanismo de transferência.

O fenômeno de transferência horizontal de genes não está limitado somente aos procariotos. Tanto os eucariotos quanto os procariotos são quimeras. Os genes informacionais dos eucariotos são derivados principalmente de um organismo relacionado com o *Methanococcus*, e seus genes operacionais são derivados fundamentalmente de proteobactérias, com algumas contribuições de cianobactérias e metanobactérias. Quase todos os genes informacionais de *Methanococcus* são semelhantes aos de levedura. Pelo menos oito genes humanos são observados no genoma de *M. tuberculosis*.

As observações apontam para o modelo de um "organismo universal", um mercado genético comum, ou até mesmo uma rede mundial de DNA (*World Wide DNA Web*), da qual os organismos podem obter genes à vontade! Como isso pode ser conciliado com o fato de que a distinção entre as espécies foi mantida? A explicação convencional estabelece que o mundo vivo contenha "nichos" ecológicos aos quais espécies próprias estão adaptadas. É o caráter discreto (composto de partes distintas) dos nichos que explica o caráter discreto (composto de indivíduos distintos) das espécies. Esta explicação, no entanto, depende da estabilidade da hereditariedade natural para manter a adaptabilidade das espécies. Por que um organismo universal não seria capaz de ultrapassar os limites entre as espécies, da mesma forma que o acesso global à cultura *pop* ameaça romper os limites entre as heranças culturais nacionais e étnicas? Talvez a resposta esteja nos genes informacionais, que parecem estar menos sujeitos à transferência horizontal, e que determinam a identidade das espécies.

É interessante destacar que, embora a evidência da importância da transferência horizontal de genes seja irresistível, por muito tempo ela foi considerada rara e insignificante. A razão deste desconforto intelectual é clara: a transmissão de genes de pais para filhos está no cerne do modelo darwiniano de evolução biológica, onde a seleção (reprodução diferenciada) de fenótipos parentais altera a freqüência gênica na próxima geração. O fato de a prole ser capaz de adquirir genes de outras fontes que não os seus pais "dá um sabor" de lamarckismo e de alternativas desacreditadas para o modelo evolutivo. A árvore evolucionária como princípio da organização das relações biológicas é um conceito fortemente enraizado: os cientistas demonstram seu fervor ambientalista no seu comprometimento com essas árvores, mesmo quando elas não constituem um modelo adequado para uma rede de parentescos. Talvez seja necessário recordar que Darwin não tinha conhecimento dos genes, e que o mecanismo que gerou a variação sobre a qual a seleção poderia agir era um mistério para ele. Talvez ele tivesse aceitado a transferência horizontal de genes com mais facilidade do que os seus discípulos!

Genômica comparativa de eucariotos

A comparação dos genomas de levedura, mosca, verme e do homem revelou 1.308 grupos de proteínas que são comuns aos quatro organismos. Esses grupos formam um conjunto central de proteínas conservadas para funções básicas, incluindo metabolismo, replicação e reparo do DNA e tradução.

Essas proteínas são compostas por domínios protéicos individuais, incluindo proteínas de um único domínio, proteínas oligoméricas e proteínas modulares contendo muitos domínios (a maior delas, a proteína muscular titina, contém entre 250 e 300 domínios). As proteínas do verme e da mosca apresentam cerca de três vezes mais domínios do que as proteínas de leveduras. As proteínas humanas são compostas do dobro de domínios que as proteínas de verme e mosca. A maioria desses domínios também é observada em bactérias e arqueobactérias, mas alguns são específicos de (e provavelmente originados em) vertebrados (ver Tabela). Essas proteínas incluem as que medeiam atividades únicas aos organismos vertebrados, como as proteínas de defesa e do sistema imune e as proteínas do sistema nervoso; apenas uma delas é uma enzima, uma ribonuclease.

Distribuição de prováveis homólogas de proteínas humanas preditas

Apenas vertebrados	22%
Vertebrados e outros animais	24%
Animais e outros eucariotos	32%
Eucariotos e procariotos	21%
Sem homólogos em animais	1%
Apenas procariotos	1%

A criação de novas proteínas ou novos domínios é um evento raro. A criação de diferentes combinações dos domínios existentes em níveis crescentes de complexidade é muito mais comum. Um mecanismo comum é o acréscimo de domínios nas terminações das proteínas modulares (ver Figura 2.3). Este processo pode ocorrer de maneira independente e tomar direções distintas em diferentes filos.

A duplicação gênica seguida de divergência é um mecanismo de criação de famílias de proteínas. Por exemplo, existem 906 genes e pseudogenes para receptores olfativos no genoma humano. Estima-se que esses receptores se liguem a cerca de 10.000 moléculas odoríferas. Foram encontrados homólogos em levedura e outros fungos (algumas comparações *são* odoríferas), mas é a necessidade dos vertebrados de terem um olfato altamente desenvolvido que levou esta família de genes a este nível de multiplicação e especialização. Oitenta por cento dos genes para receptores olfativos estão organizados em grupos. Compare-os com o pequeno tamanho dos grupos de genes da globina (página 110), proteína para a qual não houve necessidade de tamanha variedade.

Recursos da *web*: Bancos de dados de genomas

Listas de genomas completos:

`http://www.ebi.ac.uk/genomes/`
`http://pir.georgetown.edu/pirwww/search`
`http://www.nslij-genetics.org/seq/`

Bancos de dados específicos para cada organismo:

`http://fugu.biology.qmul.ac.uk/`
`http://www.litbio.org/`
`http://www.bioinformatik.de/cgi-bin/browse/Catalog/Databases/Genome_Projects/`

Figura 2.3 Evolução pelo acréscimo de domínios, de moléculas relacionadas com a *peregrin*, uma proteína humana com uma provável função na regulação da transcrição. A homóloga de *C. elegans*, lin-49, é essencial para o desenvolvimento normal do verme. A homóloga de levedura é um componente do complexo histona acetiltransferase. As proteínas contêm os seguintes domínios: ZNF = dedo de zinco tipo C_2H_2 (não confunda com o acetileno; C e H indicam cisteína e histidina, respectivamente); EP1 e EP2 = *enhancer polycomb* 1 e 2; PHD = homeodomínio de plantas, um domínio repressor contendo o dedo de zinco tipo $C_4H_3C_3$; BR = domínio bromo; PWWP = domínio que contém o motivo de seqüência Pro-Trp-Trp-Pro.

Leituras recomendadas

C. elegans sequencing consortium (1999), How the worm was won. The *C. elegans* genome sequencing project, *Trends Genet.*, **15**, 51-58. [Descrição do projeto no qual o seqüenciamento rápido e em larga escala de DNA foi desenvolvido originalmente e dos resultados; o primeiro genoma de um metazoário a ser seqüenciado.]

Ureta-Vidal, A., Ettwiller, L. & Birney, E. (2003), Comparative genomics: genome-wide analysis in metazoan eukaryotes, *Nature Reviews Genetics*, **4**, 251-262. [Uma introdução ao repertório de genomas seqüenciados e como genomas de diferentes espécies iluminam uns aos outros.]

Doolittle, W. F. (1999), Lateral genomics, *Trends Cell Biol.*, **9**, M5-M8. [Discute como a transferência horizontal incomoda as idéias tradicionais sobre evolução.]

Koonin, E. V. (2000), How many genes can make a cell: The minimal-gene-set concept, *Annu. Rev. Genomics Hum. Genet.*, **1**, 99-116. [Um resumo de trabalhos sobre genômica comparativa.]

Kwok, P. -Y. & Gu, Z. (1999), SNP libraries: why and how are we building them? *Molecular Medicine Today*, **5**, 538-543. [Discute o progresso e a base lógica dos bancos de dados de polimorfismos de um único nucleotídeo.]

Publicações dos rascunhos das seqüências do genoma humano apareceram em fascículos especiais da *Nature*, em 15 de fevereiro de 2001, contendo os resultados do Projeto Genoma Humano financiado publicamente, e da *Science*, em 16 de fevereiro de 2001, contendo os resultados produzidos pela empresa Celera Genomics. Estes fascículos são um marco da ciência.

O periódico *Genome Research*, Volume **11**, Número 5, publicado em maio de 2001, é dedicado ao genoma humano.

A conclusão do genoma humano em 2003 foi anunciada em meios de publicidade e descrita em volumes das revistas *Nature* e *Science*:

Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. (2003), A vision for the future of genomics research, *Nature*, **422**, 835-847.

The ENCODE Project Consortium (2004), The ENCODE (ENCyclopedia of DNA Elements) Project, Science, **306**, 636-640. [Seqüenciamento comparativo sistemático.]

Building on the DNA Revolution, Special section, *Science*, **300**, 11 Apr. 2003, p. 277 ff.

International Human Genome Sequencing Consortium (2004), Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931-945. [O final do início, mais precisamente.]

Southan, C. (2004), Has the yo-yo stopped? An assessment of human protein-coding gene number, *Proteomics*, **4**, 1712-1726. [Quantos genes há no genoma humano? As estimativas estão sempre mudando.]

Bentley, D. R. (2004), Genomes for medicine, *Nature*, **429**, 440-445. [Resumo e discussão sobre os resultados de seqüenciamentos genômicos e suas aplicações.]

Exercícios, problemas e *weblemas*

Exercícios

2.1 A composição geral de bases do genoma de *E. coli* é A = T = 49,2%, G = C = 50,8%. Em uma seqüência aleatória de 4.639.221 nucleotídeos com estas proporções, qual é o número esperado de ocorrências da seqüência CTAG?

2.2 O genoma de *E. coli* contém um número de pares de enzimas que catalisam as mesmas reações. Como isso afetaria os experimentos nocaute (deleção ou inativação de genes individuais) para o discernimento de funções?

2.3 Quais das categorias utilizadas para classificar as funções das proteínas de levedura (ver página 112) seriam apropriadas para classificar proteínas de um genoma procariótico?

2.4 O que ocorreu primeiro, a aterrissagem do homem na Lua ou a descoberta das fendas hidrotérmicas no fundo do mar? Primeiro arrisque um palpite, depois verifique as datas.

2.5 A síndrome de Gardner é uma condição na qual um grande número de pólipos se desenvolve na porção inferior do trato gastrintestinal, levando inevitavelmente ao câncer se não tratada. Em todos os casos observados, um dos pais do paciente também apresenta os sintomas. Qual é o padrão de herança desta doença?

2.6 O gene do retinoblastoma é transmitido junto com o gene da esterase D, ao qual está ligado. No entanto, qualquer um dos alelos da esterase D pode ser transmitido com qualquer um dos alelos do retinoblastoma. Como saber se o retinoblastoma não é um efeito direto do fenótipo esterase D?

2.7 Se todas as células somáticas de um organismo têm a mesma seqüência de DNA, por que é necessário construir bibliotecas de cDNA de diferentes tecidos?

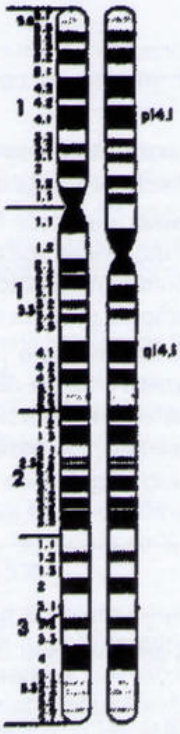
2.8 Suponha que você esteja tentando identificar um gene que causa uma doença humana. Você encontrou um marcador genético a 0,75 cM do gene da doença. Dentro de aproximadamente quantos pares de base você localizou o gene que está procurando? Quantos genes se espera que esta região possa conter?

2.9 A neuropatia óptica hereditária de Leber (LHON) é uma condição hereditária que pode causar a perda da visão central, resultado de mutações no DNA mitocondrial. Você é solicitado a aconselhar uma mulher que possui mitocôndrias normais e um homem com LHON, que pretendem se casar. Que conselho você daria a este casal sobre os riscos de seus filhos desenvolverem LHON?

2.10 A deficiência de glicose-6-fosfato desidrogenase é uma mutação recessiva, em um único gene, ligada ao cromossomo X e que afeta centenas de milhares de pessoas. As conseqüências clínicas incluem anemia hemolítica e icterícia neonatal persistente. Esse gene ainda não foi eliminado da população, pois confere resistência à malária. Neste caso, o conhecimento geral de vias metabólicas tornou capaz a identificação da proteína responsável pela doença. Dada a seqüência de aminoácidos da proteína, como você determinaria a localização do gene correspondente no cromossomo?

2.11 Antes de o DNA ser reconhecido como sendo o material genético, a natureza do gene – em termos bioquímicos detalhados – era obscura. Na década de 1940, G. Beadle e E. Tatum observaram que mutações em uma única base podiam bloquear etapas específicas em vias bioquímicas. Com base nesta observação, eles propuseram a hipótese de *um gene – uma enzima*. Em uma fotocópia da Figura 2.1, desenhe linhas que unam os genes da figura às etapas numeradas da seqüência de reações na via. Até que ponto os genes do óperon trp satisfazem a hipótese de um gene – uma enzima? E até que ponto eles apresentam exceções?

2.12 A figura mostra o cromossomo 5 humano (esquerda) e o cromossomo correspondente de chimpanzé. Em uma fotocópia desta figura, indique quais regiões apresentam inversões do padrão de bandeamento.



(De: Yunis, J.J., Sawyer, J.R. & Dunham, K. (1980), The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee, *Science*, **208**, 1145-1148.)

2.13 Descreva, em termos gerais, como a figura FISH apresentada na Gravura II seria se a região afetada do cromossomo 20 não tivesse sofrido deleção, mas translocação para outro cromossomo.

2.14 Foram inseridas 755 ORFs no genoma de *E. coli* por transferência horizontal desde a sua divergência de *Salmonella* a 14,4 milhões de anos atrás. Qual é a taxa média de transferência horizontal em kpb/ano? A quantas proteínas típicas (de aproximadamente 300 aminoácidos) esse valor médio corresponderia? Qual é a porcentagem de genes conhecidos que foram inseridos no genoma de *E. coli* por transferência horizontal?

2.15 Até que ponto o genoma de um ser vivo pode ser comparado a um banco de dados? Quais das seguintes características são compartilhadas por genomas de seres vivos e bancos de dados computacionais? Quais das seguintes propriedades são exclusivas dos genomas de seres vivos? Quais das seguintes características são exclusivas dos bancos de dados computacionais?

(a) Servem como repositório da informação.

(b) São auto-interpretativos.

(c) Cópias diferentes não são idênticas.

(d) Cientistas podem detectar erros.

(e) Cientistas podem corrigir erros.

(f) Existe responsabilidade planejada e organizada para reunir e disseminar a informação.

Problemas

2.1 Qual é a evidência experimental que mostra que o mapa de ligação genética de qualquer cromossomo é ordenado de forma linear?

2.2 Para *M. genitalium* e *H. influenzae*, quais são os valores (a) da densidade gênica em genes/kpb, (b) do tamanho médio dos genes em pb, (c) do número de genes? Qual fator contribui predominantemente para a redução do tamanho do genoma de *M. genitalium* em relação ao genoma de *H. influenzae*?

2.3 Estima-se que o sistema imune humano pode produzir 10^{15} anticorpos. Seria possível que, para este número enorme de proteínas, cada uma fosse codificada por um gene específico, com a diversidade surgindo por eventos de duplicação e divergência? Um gene típico para uma molécula de IgG possui cerca de 2.000 pb.

Weblemas

2.1 Em fotocópias das Figuras 1.2, 1.3 e 1.4, indique a posição das espécies cujas seqüências completas dos genomas são conhecidas (`http://www.ebi.ac.uk/genomes/`).

2.2 Quais são as diferenças entre o código genético padrão e o código genético das mitocôndrias dos vertebrados?

2.3 Qual é a localização cromossômica do gene humano da mioglobina?

2.4 Qual é o número de ocorrências do tetrapeptídeo CTAG no genoma de *E. coli*? A sua ocorrência é maior ou menor do que o esperado para uma seqüência aleatória de mesmo comprimento e composição nucleotídica que o genoma de *E. coli*?

2.5 Faça um histograma do número cumulativo das seqüências de genomas completas para cada ano a partir de 1995.

2.6 (a) Quantas ORFs preditas existem no cromossomo X de levedura? (b) Quantos são os genes para tRNAs?

2.7 Qual aminoácido está totalmente ausente nas proteínas de *M. genitalium*? Como o código genético de *M. genitalium* difere do código-padrão?

2.8 No genoma humano, 1 cM equivale a ~10^6 pb. Em levedura, aproximadamente quantos pares de base correspondem a 1 cM?

2.9 Espermatozóides são células nadadoras ativas e contêm mitocôndrias. Na fertilização, o conteúdo integral do espermatozóide entra no óvulo. Como então todo o DNA mitocondrial é herdado apenas da mãe?

2.10 O quadro nas páginas 110 e 111 mostra as duplicações e divergências que originaram os grupos de genes para a globina α e β humanas atuais. (a) Em qual espécie, proximamente relacionada aos ancestrais dos humanos, essas divergências ocorreram? (b) Em qual espécie, relacionada aos ancestrais dos humanos, o padrão de expressão ligado ao desenvolvimento ($\zeta_2\epsilon_2$ = embrionário; $\alpha_2\gamma_2$ = fetal; $\alpha_2\beta_2$ = adulto) surgiu?

2.11 Os grupos de línguas estão mais correlacionados com variações na seqüência do DNA mitocondrial ou do cromossomo Y? Sugira uma explicação para o resultado observado.

2.12 A mutação que causa a anemia falciforme é a alteração de uma única base, A→T, causando a alteração do tipo Glu→Val na posição 6 da cadeia β da hemoglobina. A alteração da base ocorre na seqüência 5'-GTGAG-3' (*normal*) → 5'-GTGTG-3' (*mutante*). Qual enzima de restrição é utilizada para distinguir essas seqüências, a fim de detectar portadores? Qual é a especificidade dessa enzima?

2.13 Qual mutação é a causa mais comum da fenilcetonúria (PKU)?

2.14 Encontre três exemplos de mutações no gene CFTR (associado com a fibrose cística) que reduzam, mas não eliminem totalmente, a função dos canais de íons cloreto. Quais são os sintomas clínicos dessas mutações?

2.15 Encontre um exemplo de doença genética que seja (a) autossômica dominante, (b) autossômica recessiva (além da fibrose cística), (c) dominante ligada ao X, (d) recessiva ligada ao X, (e) ligada ao Y, (f) o resultado de anomalias no DNA mitocondrial (além de LHON).

2.16 (a) Identifique um estado norte-americano onde os recém-nascidos são testados para a homocisteinúria rotineiramente. (b) Identifique um estado norte-americano onde os recém-nascidos não são testados para a homocisteinúria. (c) Identifique um estado norte-americano onde os recém-nascidos são testados para a biotinidase rotineiramente. (d) Identifique um estado norte-americano onde os recém-nascidos não são testados para a biotinidase. (e) Quais são as conseqüências clínicas da não-detecção da homocisteinúria e da deficiência de biotinidase?

2.17 (a) Qual é a função normal da proteína que é defectiva na doença de Menke? (b) Existe um homólogo deste gene no genoma de *A. thaliana*? (c) Se existir, qual é a função desse gene em *A. thaliana*?

2.18 A *distrofia muscular de Duchenne* (DMD) é uma doença hereditária ligada ao X que causa fraqueza muscular progressiva. Pacientes com DMD geralmente perdem a capacidade de andar aos 12 anos de idade, e a expectativa de vida é de 20 a 25 anos. A *distrofia muscular de Becker* (BMD) é uma condição menos grave

envolvendo o mesmo gene. As duas condições são causadas pela deleção de um único gene, a distrofina. Na DMD, ocorre a ausência completa da proteína funcional e na BMD observa-se uma forma truncada da proteína que retém alguma funcionalidade. Algumas das deleções que causam DMD são mais longas do que outras que causam BMD. O que distingue as duas classes de deleções que originam estas duas síndromes?

2.19 Qual cromossomo da vaca contém uma região homóloga à região 8q21.12 do cromossomo humano?

CAPÍTULO 3

Bancos de Dados e Recuperação de Informação

Sumário do capítulo

Introdução 138

Indexação de bancos de dados e especificação de termos de pesquisa 138

Monitoração dos resultados da pesquisa 140

Análise dos dados recuperados 141

Os bancos de dados 141

Bancos de dados de seqüências de ácidos nucléicos 142

Bancos de dados de genomas 144

Bancos de dados de seqüências de proteínas 144

Bancos de dados de estruturas 148

Bancos de dados especializados ou "boutique" 155

Bancos de dados de padrões de expressão e de proteômica 156

Bancos de dados de vias metabólicas 158

Bancos de dados bibliográficos 159

Avaliações sobre bancos de dados e servidores de biologia molecular 159

Acessos a bancos de dados 160

Acesso a bancos de dados de biologia molecular 161

ENTREZ 161

O Sequence Retrieval System (SRS) 168

O Protein Information Resource (PIR) 169

ExPASy – Expert Protein Analysis System 170

Ensembl 171

Para onde vamos a partir daqui? 172

Leituras recomendadas 172

Exercícios, problemas e weblemas 173

Objetivos do aprendizado

1. Entender os tipos de dados gerais relacionados a moléculas e processos da vida que são agrupados para dar suporte à pesquisa e a aplicações em biologia, medicina, agricultura e tecnologia.
2. Conhecer a infra-estrutura básica da bioinformática, em termos dos sítios na *web* e das responsabilidades dos principais projetos de bancos de dados.
3. Entender os conceitos básicos de recuperação de informação, incluindo como formular consultas a bancos de dados.
4. Desenvolver habilidades com os instrumentos de pesquisa gerais na *web* e com sítios específicos de bioinformática.
5. Saber como pesquisar por informações específicas sobre seqüências, estruturas, vias metabólicas, relações com doenças, e como realizar análises sobre os dados recuperados.

Introdução

Este capítulo apresenta as habilidades necessárias para se realizar recuperação de informação. Essas habilidades lhe permitirão fazer um uso eficiente dos bancos de dados. O objetivo é familiarizá-lo com as operações básicas. Dessa forma, você terá maior facilidade para melhorar e desenvolver sua técnica. Na verdade, muitos bancos de dados disponibilizam tutoriais que facilitam a exploração dos seus recursos.

Indexação de bancos de dados e especificação de termos de pesquisa

Um índice é um conjunto de ponteiros para informações em um banco de dados. Ao pesquisar toda a *World Wide Web*, ou um banco de dados especializado em biologia molecular, você propõe um ou mais termos de pesquisa e um programa procura por eles em suas tabelas de índices. O modelo consiste no seguinte: todo banco de dados é composto de entradas – pacotes discretos e coerentes de informação. O *software* de recuperação de informação identifica entradas com conteúdos relevantes ao seu interesse. Um exemplo do paradigma mais simples é aquele em que você submete o termo "cavalo" e o programa retorna uma lista de entradas que contêm o termo "cavalo".

Uma pesquisa completa na *web* retornaria informações sobre muitos aspectos diferentes de cavalos – biologia molecular, raça, corrida, poemas sobre cavalos – a maioria dos quais não lhe interessa. Para uma pesquisa bem-sucedida, não é suficiente mencionar o que você *quer* – você deve assegurar que as respostas desejadas não fiquem encobertas por resultados insignificantes, ou lixo. (Claro que lixo é simplesmente qualquer informação que não lhe interessa.)

Para direcionar os resultados, ferramentas de recuperação de informação aceitam múltiplos termos de consulta ou palavras-chave. Uma procura por "álcool desidrogenase de fígado de cavalo" produziria respostas específicas para essa enzima. A pesquisa identificaria entradas que contêm todas as quatro palavras-chave que você submeteu: álcool AND desidrogenase AND fígado AND ca-

valo.* A pesquisa não retornaria poemas sobre cavalos entre seus primeiros resultados (exceto no caso improvável de um poema conter todas essas quatro palavras-chave).

É possível realizar pesquisas utilizando outras combinações lógicas de termos de indexação. Por exemplo, se uma ferramenta de busca não tivesse conhecimento sobre diferenças de escrita entre o inglês britânico e o americano, seria necessário que ela fosse capaz de fazer uma pesquisa por "hemoglobin OR haemoglobin".** (Note que a pesquisa por "hemoglobin haemoglobin" provavelmente seria interpretada como "hemoglobin AND haemoglobin", a qual resultaria somente em documentos escritos por comitês internacionais ou expatriados com dificuldades ortográficas.)

Se você quisesse saber algo sobre outras desidrogenases, você poderia procurar por "desidrogenase NOT álcool".*** Isso retornaria entradas que contivessem o termo desidrogenase, mas NÃO a palavra álcool. Você encontraria entradas sobre lactato desidrogenase, malato desidrogenase, etc. Você perderia referências a artigos de revisão que comparassem álcool desidrogenases com outras desidrogenases, ou alinhamentos das seqüências de muitas desidrogenases incluindo a álcool desidrogenase. Você poderia se arrepender dessas perdas.

Muitos instrumentos de pesquisa em bancos de dados permitirão expressões lógicas complexas, tais como "(haemoglobin OR hemoglobin) AND (dehydrogenase NOT alcohol)". A construção dessas expressões é um exercício de teoria de conjuntos, e pode ser auxiliada pelo desenho de diagramas de Venn. Embora a lógica de uma pesquisa seja independente do *software* utilizado para consultar um banco de dados, programas diferentes exigem sintaxes diferentes para expressar as mesmas condições. Por exemplo, a consulta por "dehydrogenase NOT alcohol" poderia ser realizada como "DEHYDROGENASE – ALCOHOL" ou "DEHYDROGENASE! ALCOHOL".

Bancos de dados especializados, incluindo os de biologia molecular, impõem uma estrutura particular para a informação com o objetivo de separá-la em diferentes categorias. Isso é essencial. Existem, atualmente, cientistas biomédicos em atividade chamados E(lisabetta) Coli, (John D.) Yeast, (Patrice) Rat, e um grande número de Rabbits, assim como muitos Crystals e Blots. Se você quisesse encontrar artigos publicados por esses pesquisadores, seria ingênuo efetuar uma pesquisa geral em bancos de dados de biologia molecular com quaisquer desses sobrenomes. Muitos bancos de dados fornecem indexação e pesquisas separadas para diferentes categorias de informação. Eles permitem pesquisas por artigos nos quais E. Coli é um AUTHOR.

Algumas dessas categorias, como, por exemplo, taxonomia, têm **vocabulários controlados**. Geralmente, nesses casos, o vocabulário é apresentado ao usuário na forma de menus suspensos. Para se fazer uma pesquisa por "globin NOT mammal" e obter relativamente poucas entradas sobre globinas que não pertencem a mamíferos, ao invés de muitas entradas sobre globinas, incluindo hemoglobinas humanas, que não mencionam explicitamente o termo mamífero, é necessário um sistema de recuperação de informação que "entenda" a hierarquia taxonômica. Vocabulários controlados – conjunto limitado, explícito

* N. de T. AND é o termo lógico da álgebra booleana e representa a conjunção inclusiva "E" em consultas a bancos de dados.

** N. de T. OR é o termo lógico da álgebra booleana e representa a conjunção exclusiva "OU" em consultas a bancos de dados.

*** N. de T. NOT equivale ao termo lógico de consulta para a negação "NÃO" na álgebra booleana.

e cuidadosamente definido de termos – são importantes também em consultas distribuídas entre diversos bancos de dados.

Um problema técnico que freqüentemente gera dificuldade é como realizar pesquisas com termos que contenham caracteres incomuns como acentos ou tremas, letras gregas e, como já mencionado, diferenças de escrita entre o inglês britânico e o americano. Um banco de dados especializado como o ENTREZ do NCBI pode tratar as diferenças de escrita britânica e americana utilizando um dicionário de sinônimos. Programas que indexam toda a *web* geralmente não fazem isso. Portanto, ignore os acentos e espere pelo melhor.

Monitoração dos resultados da pesquisa

Quando pesquisar bancos de dados, é muito improvável que você encontre exatamente o que procura na primeira rodada de busca. Muitas vezes, você precisa modificar a consulta com base nos resultados retornados inicialmente. A maioria dos *softwares* de recuperação de informação permite pesquisas consecutivas e cumulativas, com o conjunto dos termos de pesquisa e/ou relações lógicas alterados. Por outro lado, uma vez que encontre o que estava procurando, você geralmente vai querer estender sua pesquisa para encontrar assuntos relacionados. Caso encontre a seqüência de um gene, você poderá querer saber sobre genes homólogos em outros organismos. Ou se uma estrutura tridimensional da proteína correspondente está disponível. Ou, ainda, pode querer saber sobre artigos publicados acerca desse gene.

Para essas consultas auxiliares seriam necessárias conexões entre as entradas no mesmo ou em diferentes bancos de dados. Este é um exemplo especial da questão de como "navegar" em bibliotecas eletrônicas – um problema muito difícil e que é assunto de pesquisas em andamento.

Para encontrar genes homólogos você precisaria de conexões para outros itens no *mesmo* banco de dados (um banco de dados de seqüências de genes). Para encontrar estruturas, ou referências bibliográficas relacionadas ao gene, você precisaria de conexões *entre* diferentes bancos de dados (do banco de dados de seqüências de genes para bancos de dados de estruturas tridimensionais ou para um banco de dados bibliográficos). Como o número de bancos de dados cresce, a comunicação entre eles tem se tornado um objetivo de alta prioridade. De fato, a interatividade entre bancos de dados de biologia molecular está se tornando cada vez mais eficaz, de tal forma que essas operações são razoavelmente fáceis agora – antigamente uma pessoa precisava fazer pesquisas isoladas em bancos de dados separados. Esta é uma generalização do modelo original de um banco de dados, como um conjunto fechado de entradas independentes que podem ser selecionadas somente por seus conteúdos indexados.

Até certo ponto, as atividades de bancos de dados em bioinformática podem ser classificadas como *arquivamento* – cujos maiores objetivos são a conservação e a organização – e *interpretação* – a compilação de informações biológicas em uma forma mais útil para dar suporte à pesquisa. Diferentes bancos de dados se especializam em diferentes tipos de informações – seqüências de ácidos nucléicos, seqüências de proteínas, estruturas – por motivos em parte históricos e em parte por causa da necessidade de diferentes habilidades gerenciais específicas. Bancos de dados interpretativos, ou secundários, são livres para combinar informações de quaisquer fontes disponíveis. Na maioria dos casos, os projetos de arquivamento e interpretação dos dados são executados pela mesma instituição e até mesmo pelas mesmas pessoas.

Dois aspectos do desenvolvimento de bancos de dados em bioinformática são visíveis. Um é o grande crescimento de projetos de bancos de dados individuais que recombinam os dados armazenados de diferentes maneiras. O outro é a combinação de muitos bancos de dados individuais em sítios "guarda-chuva" (sítios abrangentes, do tipo portais). Realmente, não há paradoxo – ambos estão progredindo. (Genes também se multiplicam e combinam.)

A maioria das unificações de bancos de dados são meramente extensões, mas com aparência familiar, das colaborações ou dos esforços competitivos que, na maioria dos casos, já existiam anteriormente. Veremos, por exemplo, que os bancos de dados de seqüências de proteínas estão coordenando suas atividades como o UniProt; e que os bancos de dados de estruturas de proteínas estão coordenando suas atividades como o Worldwide Protein Data Bank (wwPDB). O Interpro é um banco de dados "guarda-chuva" que integra conteúdos, características e anotações de bancos de dados individuais de famílias de proteínas, domínios e sítios funcionais e contém conexões para outros bancos, incluindo a classificação funcional do Gene Ontology Consortium™. Atualmente, o Interpro inclui os bancos de dados PROSITE, Pfam, PRINTS, SMART e ProDom, e ainda pretende incorporar outros. (Resistir é inútil.)

Análise dos dados recuperados

Algumas vezes você deseja utilizar o resultado de uma pesquisa como entrada em outro programa. Por exemplo, se você identifica uma seqüência de uma proteína de interesse, poderia querer executar uma pesquisa PSI-BLAST. Este não é estritamente um problema de busca em bancos de dados. Antigamente você teria de executar os programas em separado, incluindo, manualmente, a seqüência obtida na aplicação desejada. Entretanto, como pesquisas em múltiplos bancos de dados, sistemas de recuperação de informação em biologia molecular geralmente oferecem recursos para iniciar processos como este. Isso faz com que as suas sessões de trabalho no computador sejam bem mais eficientes.

Os bancos de dados

Embora nosso conhecimento de dados de seqüências e estruturas biológicas esteja muito longe de estar completo, ele já apresenta um tamanho respeitável e cresce muito rapidamente. Muitos cientistas estão trabalhando para gerar os dados, ou para executar projetos de pesquisa analisando os resultados. O arquivamento e a distribuição desses dados são realizados por organizações mantenedoras de bancos de dados específicos.

O arquivamento de dados na área de bioinformática era originalmente realizado por grupos de pesquisa individuais, motivados pelo interesse na ciência associada àqueles dados. Como as demandas por equipamentos e pessoal aumentaram – e a natureza das habilidades necessárias mudou para incluir uma ênfase maior em computação – esse arquivamento passou a ser de responsabilidade de projetos específicos nacionais e até internacionais, em uma escala realmente muito grande. Qualquer pessoa que acompanhou a história desses projetos desde o início não pode deixar de ficar perplexa com o seu crescimento vertiginoso, de pequenos projetos com objetivos modestos e mal financiados, executados por umas poucas pessoas dedicadas, a um assunto de

extremo interesse da indústria multinacional, com a aquisição e venda de pequenas empresas do ramo.

Os bancos de dados primários, relacionados com macromoléculas biológicas, incluem:

- Seqüências de ácidos nucléicos, incluindo genomas completos
- Seqüências de aminoácidos de proteínas
- Estruturas de proteínas e ácidos nucléicos
- Estruturas cristalográficas de pequenas moléculas
- Funções de proteínas
- Padrões de expressão de genes
- Vias metabólicas e redes de interação e controle
- Publicações

Bancos de dados de seqüências de ácidos nucléicos

O arquivamento mundial de seqüências de ácidos nucléicos é uma parceria tríplice entre o National Center for Biotechnology Information (Estados Unidos), o EMBL Data Library (European Bioinformatics Institute, Reino Unido) e o DNA Data Bank of Japan (National Institute of Genetics, Japão). Esses grupos trocam informações diariamente. Como resultado, os dados brutos são idênticos, embora o formato em que são armazenados e a natureza da anotação variem entre eles. Esses bancos de dados organizam, arquivam e distribuem seqüências de DNA e RNA coletadas de projetos genoma, publicações científicas e depósitos de patentes. Para garantir que esses dados fundamentais permaneçam disponíveis livremente, revistas científicas exigem o depósito de novas seqüências de nucleotídeos em um banco de dados como uma condição para a publicação de um artigo. Condições semelhantes se aplicam a seqüências de aminoácidos e a estruturas de ácidos nucléicos e proteínas.

Os bancos de dados de ácidos nucléicos, como distribuídos, são coleções de registros ou entradas. Cada entrada tem a forma de um arquivo texto contendo dados e anotações para uma seqüência contígua única. Muitas entradas são reunidas a partir de diversos artigos publicados que descrevem fragmentos sobrepostos de uma seqüência completa. Outras são genomas completos.

As entradas têm um ciclo de vida no banco de dados. Devido ao desejo de parte da comunidade de usuários de poder ter rápido acesso aos dados, novas entradas são disponibilizadas antes que a anotação esteja completa e que verificações sejam realizadas. As entradas passam pelas seguintes classes até chegarem ao formato final:

$$\text{Não anotada} \rightarrow \text{Preliminar} \rightarrow \text{Não revisada} \rightarrow \text{Padrão}$$

Raramente uma entrada "morre" – algumas foram removidas quando se determinou que estavam incorretas.

Uma amostra de uma entrada de seqüência de DNA do EMBL Data Library, incluindo anotações assim como os dados da seqüência, é o gene do inibidor da tripsina pancreática bovina (o Quadro mostra parte dessa entrada, omitindo a maior parte da seqüência propriamente dita).

A entrada no EMBL Data Library para o gene do inibidor da tripsina pancreática bovina

```
ID   BTBPTIG    standard; DNA; MAM; 3998 BP.
XX
AC   X03365; K00966;
XX
DT   18-NOV-1986 (Rel. 10, Created)
DT   20-MAY-1992 (Rel. 31, Last updated, Version 3)
XX
DE   Bovine pancreatic trypsin inhibitor (BPTI) gene
XX
KW   Alu-like repetitive sequence; protease inhibitor;
KW   trypsin inhibitor.
XX
OS   Bos taurus (cattle)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC   Theria; Eutheria; Artiodactyla; Ruminantia; Pecora; Bovidae.
XX
RN   [1]
RP   1-3998
RA   Kingston I.B., Anderson S.;
RT   "Sequences encoding two trypsin inhibitors occur in strikingly
RT   similar genomic environments";
RL   Biochem. J. 233:443-450(1986).
XX
RN   [2]
RA   Anderson S., Kingston I.B.;
RT   "Isolation of a genomic clone for bovine pancreatic trypsin
RT   inhibitor by using a unique-sequence synthetic dna probe";
RL   Proc. Natl. Acad. Sci. U.S.A. 80:6838-6842(1983).
XX
DR   SWISS-PROT; P00974; BPT1_BOVIN.
XX
CC   Data kindly reviewed (08-DEC-1987) by Kingston I.B.
XX
FH   Key             Location/Qualifiers
FH
FT   misc_feature    795..800
FT                   /note="pot. polyA signal"
FT   misc_feature    835..839
FT                   /note="pot. polyA signal"
FT   repeat_region   837..847
FT                   /note="direct repeat"
FT   misc_feature    930..945
FT                   /note="sequence homologous to Alu-like
FT                   consensus seq."
FT   repeat_region   1035..1045
FT                   /note="direct repeat"
FT   misc_feature    2456..2461
FT                   /note="pot. splice signal"
FT   CDS             2470..2736
FT                   /note="put. precursor"
FT   misc_feature    2488..2489
FT                   /note="pot. intron/exon splice junction"
FT   misc_feature    2506..2507
FT                   /note="pot. intron/exon splice junction"
FT   CDS             2512..2685
FT                   /note="trypsin inhibitor (aa 1-58)"
FT   misc_feature    2698..2699
FT                   /note="pot. exon/intron splice junction"
FT   misc_feature    3690..3695
FT                   /note="pot. polyA signal"
FT   misc_feature    3729..3733
FT                   /note="pot. polyA signal"
XX
SQ   Sequence 3998 BP; 1053 A; 902 C; 892 G; 1151 T; 0 other;
     aattctgata atgcagagaa ctggtaagga gttctgattg ttctgcttga ttaaatgggt
     tgtaacagga tagtgtcttg tcctgatcct agcattcata tggtgtgtgt tctggggcaa
     gtcatctgca gtttcttcac ctgaacaggg ggaccaggtt acatgagttt cttaaaagat
     taccagtcat gagtatgaag agtttacact ttcctgatca atgacgtcca tttcccatca

                        3.720 nucleotídeos foram removidos ...

     gccaggtcaa actttggggt gtgttatttc cctgaatt
//
```

Uma tabela de características (linhas que começam por FT, de Feature Tables, em inglês) é um componente da anotação de uma entrada que descreve propriedades de regiões específicas, por exemplo, seqüências codificadoras (CDS). Visto que essas seqüências são planejadas para serem lidas por pro-

gramas de computador – por exemplo, programas que traduzem uma região codificadora em uma seqüência de aminoácidos – elas têm um formato cuidadosamente controlado e um vocabulário mais restrito. O desenvolvimento de vocabulários controlados e de um dicionário-padrão e de sinônimos compartilhados, para palavras-chave e tabelas de características, é importante também para estabelecer conexões entre os diferentes bancos de dados.

A tabela de características pode indicar regiões que:

- executam ou afetam a função
- interagem com outras moléculas
- afetam a replicação
- estão envolvidas em recombinação
- são uma unidade repetida
- possuem estrutura secundária ou terciária
- são revisadas ou corrigidas

Bancos de dados de genomas

Embora as seqüências de genomas constituam entradas nos arquivos-padrão de seqüência de ácidos nucléicos, muitas espécies têm bancos de dados especiais que juntam a seqüência do genoma e sua anotação com outros dados relacionados às espécies.

Bancos de dados de seqüências de proteínas

Em 2002, três bancos de dados de proteínas – o Protein Information Resource no National Biomedical Research Foundation da Georgetown University Medical Center em Washington, DC, Estados Unidos; o SWISS-PROT e o TrEMBL do Swiss Institute of Bioinformatics, Genebra, e o European Bioinformatics Institute em Hinxton, Reino Unido – coordenaram seus esforços para formar o consórcio *UniProt*. Os parceiros desta iniciativa compartilham o banco de dados, mas continuam a oferecer ferramentas separadas de acesso à recuperação de informação.

O PIR originou-se do primeiro banco de dados de seqüências, desenvolvido por Margaret O. Dayhoff – a pioneira da área de bioinformática. O SWISS-PROT foi desenvolvido no Swiss Institute of Bioinformatics. O TrEMBL contém as traduções dos genes identificados nas seqüências de DNA do EMBL Data Library. Entradas do TrEMBL são consideradas como preliminares, e são convertidas – após organização e anotações extensas – em entradas completas do SWISS-PROT.

Hoje, quase todas as informações de seqüências de aminoácidos são provenientes da tradução de seqüências de ácidos nucléicos. Informações sobre ligantes, pontes dissulfeto, associações entre subunidades, modificações pós-traducionais, glicosilação, efeitos da edição do mRNA, etc. não podem ser obtidas das seqüências de genes. Por exemplo, apenas a partir da informação genética não se saberia que a insulina humana é um dímero estabilizado por pontes dissulfeto. Bancos de dados de seqüências de proteínas coletam essas informações adicionais da literatura e as disponibilizam como anotações relevantes.

Proveniente do UniProt, a entrada para a seqüência de aminoácidos do gene inibidor da tripsina pancreática bovina, no formato SWISS-PROT, é mostrada no Quadro das páginas 146-147. (A comparação do formato SWISS-PROT com os formatos do ENTREZ e PIR é o assunto do *Weblema* 3.12.)

Bancos de dados associados com o SWISS-PROT

O ENZYME DB e o PROSITE – uma coleção de motivos de seqüências – são dois bancos de dados relacionados e intimamente associados com o SWISS-PROT.

O ENZYME DB armazena as seguintes informações sobre enzimas:

- Número EC: um identificador numérico designado pela Comissão de Enzimas (autorizada pela União Internacional de Bioquímica e Biologia Molecular – IUBMB; ver `http://www.chem.qmw.ac.uk/iubmb/enzyme/`)
- Nome recomendado
- Nomes alternativos, se houver algum
- Atividade catalítica
- Co-fatores, se houver algum
- Ponteiros para o SWISS-PROT e outros bancos de dados
- Ponteiros para doenças associadas com a deficiência da enzima, se houver alguma conhecida

Uma amostra de uma entrada do banco de dados ENZYME DB

```
ID 1.14.17.3
DE PEPTIDYLGLYCINE MONOOXYGENASE.
AN PEPTIDYL ALPHA-AMIDATING ENZYME.
AN PEPTIDYLGLYCINE 2-HYDROXYLASE.
CA PEPTIDYLGLYCINE + ASCORBATE + O(2) = PEPTIDYL(2-HYDROXYGLYCINE) +
CA DEHYDROASCORBATE + H(2)O.
CF COPPER.
CC -!- PEPTIDYLGLYCINES WITH A NEUTRAL AMINO ACID RESIDUE IN THE
CC PENULTIMATE POSITION ARE THE BEST SUBSTRATES FOR THE ENZYME.
CC -!- THE ENZYME ALSO CATALYZES THE DISMUTATATION OF THE PRODUCT TO
CC GLYOXYLATE AND THE CORRESPONDING DESGLYCINE PEPTIDE AMIDE.
DR P10731, AMD_BOVIN ; P19021, AMD_HUMAN ; P14925, AMD_RAT ;
DR P08478, AMD1_XENLA; P12890, AMD2_XENLA;
```

Os dois primeiros caracteres de cada linha identificam a informação que a linha contém. Por exemplo, ID = Identificação (*Identification*), DE = Descrição (*Description*) = Nome oficial, AN = Nome(s) alternativo(s) (*Alternative name(s)*), CA= Atividade catalítica (*Catalytic Activity*), CF = Co-fator(es) (*Cofactor(s)*), CC = Comentários (*Comments*), DR = Referência no banco de dados (*Database reference*, para o SWISS-PROT).

O PROSITE contém padrões de resíduos comuns de grupos de proteínas. Este padrão (ou motivo, ou assinatura, ou impressão digital, ou molde) geralmente aparece em uma família de proteínas relacionadas por causa das exigências de sítios de ligação que restringem a evolução dessa família de proteínas. Muitas vezes, eles indicam relações distantes não detectáveis pela comparação direta de seqüências. O padrão de consenso da pirofosfatase inorgânica é: `D-[SGN]-D-[PE]-[LIVM]-D-[LIVMGC]`. Os três aspartatos (D) conservados se ligam a cátions metálicos bivalentes.

O PIR e os bancos de dados associados

O PIR mantém muitos bancos de dados sobre proteínas:

- PIR-PSD: o principal banco de dados de seqüências de proteínas
- iProClass: fornece informações abrangentes das propriedades de proteínas presentes no SwissProt

Seqüência de aminoácidos do inibidor da tripsina pancreática bovina

NiceProt View of Swiss-Prot: P00974

Entry information

Entry name	**BPT1_BOVIN**
Primary accession number	**P00974**
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 01, July 1986
Sequence was last modified in	Release 10, March 1989
Annotations were last modified in	Release 44, June 2004

Name and origin of the protein

Protein name	**Pancreatic trypsin inhibitor [Precursor]**
Synonyms	**Basic protease inhibitor** **BPI** **BPTI** **Aprotinin**
Gene name	None
From	Bos taurus (Bovine) [TaxID: 9913]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.

References

[1] SEQUENCE FROM NUCLEIC ACID MEDLINE=87283904; PubMed=2441071; Creighton T.E., Charles I.G.; "Sequences of the genes and polypeptide precursors for two bovine protease inhibitors"; J. Mol. Biol. 194:11–22(1987).

REFERÊNCIAS ADICIONAIS FORAM REMOVIDAS

Comments

- ***FUNCTION***: Inhibits trypsin, kallikrein, chymotrypsin, and plasmin.
- ***SUBCELLULAR LOCATION***: Secreted.
- ***PHARMACEUTICAL***: Available under the name Trasylol (Mile). Used for inhibiting coagulation so as to reduce blood loss during bypass surgery.
- ***SIMILARITY***: Contains 1 BPTI/Kunitz inhibitor domain.
- ***DATABASE***: Name=Trasylol; Note=Clinical information on Trasylol; `www="http://www.trasylol.com/"`.

COMENTÁRIOS ADICIONAIS FORAM REMOVIDOS

Copyright

The Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institite of Bioinformatics and the EMBL outstation–the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See `http://www.isb-sib.ch/announce/` or send an email to license@isb-sib.ch)

Cross-references

EMBL	M20934; AAD13685.1;-.
	REFERÊNCIAS CRUZADAS ADICIONAIS AO EMBL FORAM REMOVIDAS
PIR	S00277; TIBO.
PDB	1K09; 10-JUL-02.
	REFERÊNCIAS CRUZADAS ADICIONAIS AO PDB FORAM REMOVIDAS
InterPro	IPR002223; Kunitz_BPTI.
Pfam	PF00014; Kunitz_BPTI; 1.
	Pfam graphical view of domain structure.
PRINTS	PR00759; BASICPTASE.
ProDom	PD000222; Kunitz_BPTI; 1.
	[Domain structure/List of seq. sharing at least 1 domain]
SMART	SM00131; KU; 1.
PROSITE	PS00280; BPTI_KUNITZ_1;1.
	PS50279; BPTI_KUNITZ_1; 2.
	PROSITE graphical view of domain structure.
Implicit links to	HOVERGEN; BLOCKS; ProtoNet; ProtoMap; PRESAGE; DIP; ModBase; SMR; SWISS-2DPAGE; UniRef.

Keywords
Serine protease inhibitor; Signal; Pharmaceutical; 3D-structure.

Features

Key	From	To	Length	Description
SIGNAL	1	21	21	Potential.
PROPEP	22	35	14	
CHAIN	36	93	58	Pancreatic trypsin inhibitor.
PROPEP	94	100	7	
DOMAIN	40	90	51	BPTI/Kunitz inhibitor.
SITE	50	51	2	Reactive bond for trypsin.
DISULFID	40	90		
DISULFID	49	73		
DISULFID	65	86		
HELIX	38	41	4	
STRAND	53	59	7	
TURN	60	63	4	
STRAND	64	70	7	
STRAND	80	80	1	
HELIX	83	90	8	

Sequence information

Length: **100 AA** [This is the length of the unprocessed precursor]
Molecular weight: **10903 Da** [This is the MW of the unprocessed precursor]
CRC64: **6A778A4AD763FB19** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MKMSRLCLSV ALLVLLGTLA ASTPGCDTSN QAKAQRPDFC LEPPYTGPCK ARIIRYFYNA

        70         80         90        100
         |          |          |          |
KAGLCQTFVY GGCRAKRNNF KSAEDCMRTC GGAIGPWENL
```

- PIRSF: banco de dados de anotação e similaridade; cada entrada é conectada a uma lista de seqüências similares com base em relações evolucionárias que permite a anotação de funções biológicas específicas e funções bioquímicas genéricas
- RESID: um banco de dados de modificações covalentes na estrutura de proteínas (lembre-se de que características estruturais importantes de proteínas, como pontes dissulfeto, não são inferidas das seqüências de genes e não aparecerão nos bancos de dados de seqüências de proteínas derivadas unicamente da tradução de dados genômicos)

O PIR também desenvolveu o IESA: O Integrated Environment for Sequence Analysis, um sítio para recuperação de informações e para submeter cálculos.

O servidor *web* do PIR apresenta algumas das muitas ferramentas de recuperação de informações disponíveis, que incluem:

- Recuperar entradas específicas de bancos de dados
- Alinhamento par a par de seqüências
- Alinhamento múltiplo de seqüências
- PIRSF Scan: compara uma seqüência sonda com o banco de dados PIRSF de famílias de proteínas organizadas e anotadas
- Text Search: pesquisa por texto em bancos de dados
- Peptide Search: pesquisa por pequenos peptídeos (de 3 a 30 resíduos)
- Pattern Search: reconhecimento de padrões
- Related Sequences: identificação de homologias

Bancos de dados de estruturas

Bancos de dados de estruturas arquivam, anotam e distribuem conjuntos de coordenadas atômicas. O principal banco de dados de estrutura de macromoléculas biológicas é o Protein Data Bank (PDB). Ele contém estruturas de proteínas, ácidos nucléicos e uns poucos carboidratos. Iniciado pelo falecido Walter Hamilton do Brookhaven National Laboratories, Long Island, New York, Estados Unidos, em 1971, o PDB é gerenciado atualmente pelo Research Collaboratory for Structural Bioinformatics (RCSB), uma organização distribuída com base na Rutgers University, em New Jersey; no San Diego Supercomputer Center, na Califórnia; e no National Institute of Standards and Technology, em Maryland, todos nos Estados Unidos. O sítio *web* do Protein Data Bank esta disponível em `http://www.rcsb.org`.

A *home page* do PDB possui conexões para os seus próprios arquivos de dados, para material expositivo e tutorial, incluindo anúncios curtos e o PDB Newsletter, um boletim de notícias sobre o PDB, e recursos para o depósito de novas entradas e *softwares* de pesquisa especializados para recuperação e análise de estruturas.

Recentemente, o RCSB, o Molecular Structure Database, o European Bioinformatics Institute e o Protein Data Bank of Japan formaram o Worldwide Protein Data Bank (wwPDB), com o objetivo de produzir um arquivo unificado.

O Quadro mostra parte de uma entrada do Protein Data Bank* para a estrutura da tioredoxina† de *E. coli*. As informações do arquivo incluem:

- Que proteína é o assunto da entrada, e a qual espécie ela pertence
- Quem determinou a estrutura, e referências para publicações descrevendo a determinação dessa estrutura
- Detalhes experimentais sobre a determinação da estrutura, incluindo informações relacionadas à qualidade geral do resultado, como a resolução de uma estrutura determinada por difração de raios X e estatísticas estereoquímicas
- A seqüência de aminoácidos
- Que moléculas adicionais aparecem na estrutura, incluindo co-fatores, inibidores e solvente, como moléculas de água
- Atribuições de estrutura secundária: hélice, folha
- Pontes dissulfeto
- As coordenadas atômicas

Entrada 2TRX no Protein Data Bank para a tioredoxina de *E. coli*

```
HEADER    ELECTRON TRANSPORT                      19-MAR-90   2TRX
COMPND    THIOREDOXIN
SOURCE    (ESCHERICHIA $COLI)
AUTHOR    S.K.KATTI,D.M.LE*MASTER,H.EKLUND
REVDAT   2   15-JAN-93 2TRXA    1       HEADER COMPND
REVDAT   1   15-OCT-91 2TRX     0
JRNL        AUTH    S.K.KATTI,D.M.LE*MASTER,H.EKLUND
JRNL        TITL    CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA
JRNL        TITL 2 $COLI AT 1.68 ANGSTROMS RESOLUTION
JRNL        REF     J.MOL.BIOL.                   V. 212   167 1990
JRNL        REFN    ASTM JMOBAK  UK ISSN 0022-2836                 070
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH    A.HOLMGREN,B.-*O.SODERBERG,H.EKLUND,C.-*I.BRANDEN
REMARK   1  TITL    THREE-DIMENSIONAL STRUCTURE OF ESCHERICHIA COLI
REMARK   1  TITL 2 THIOREDOXIN-*S=2= TO 2.8 ANGSTROMS RESOLUTION
REMARK   1  REF     PROC.NAT.ACAD.SCI.USA         V.  72  2305 1975
REMARK   1  REFN    ASTM PNASA6  US ISSN 0027-8424                 040
REMARK   1 REFERENCE 2
REMARK   1  AUTH    B.-*O.SODERBERG,A.HOLMGREN,C.-*I.BRANDEN
REMARK   1  TITL    STRUCTURE OF OXIDIZED THIOREDOXIN TO 4.5 ANGSTROMS
REMARK   1  TITL 2 RESOLUTION
REMARK   1  REF     J.MOL.BIOL.                   V.  90   143 1974
REMARK   1  REFN    ASTM JMOBAK  UK ISSN 0022-2836                 070
REMARK   1 REFERENCE 3
REMARK   1  AUTH    A.HOLMGREN,B.-*O.SODERBERG
REMARK   1  TITL    CRYSTALLIZATION AND PRELIMINARY CRYSTALLOGRAPHIC
REMARK   1  TITL 2 DATA FOR THIOREDOXIN FROM ESCHERICHIA $COLI B
REMARK   1  REF     J.MOL.BIOL.                   V.  54   387 1970
REMARK   1  REFN    ASTM JMOBAK  UK ISSN 0022-2836                 070
REMARK   2
REMARK   2 RESOLUTION. 1.68 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT. BY THE RESTRAINED LEAST-SQUARES PROCEDURE OF J.
REMARK   3  KONNERT AND W. HENDRICKSON AS MODIFIED BY B. FINZEL
REMARK   3  (PROGRAM *PROFFT*).  THE R VALUE IS 0.165 FOR 25969
REMARK   3  REFLECTIONS IN THE RESOLUTION RANGE 8.0 TO 1.68 ANGSTROMS
REMARK   3  WITH FOBS .GT. 3.0*SIGMA(FOBS)
REMARK   3
```

→

* Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000), The Protein Data Bank Nucleic Acids Research, **28**, 235-242.

† Katti, S. K., LeMaster, D. M. & Eklund, H. (1990), Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.*, **212**,167-184.

Entrada 2TRX no Protein Data Bank para a tioredoxina de *E. coli* (*continuação*)

```
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES (THE VALUES OF
REMARK   3      SIGMA, IN PARENTHESES, ARE THE INPUT ESTIMATED
REMARK   3      STANDARD DEVIATIONS THAT DETERMINE THE RELATIVE
REMARK   3      WEIGHTS OF THE CORRESPONDING RESTRAINTS)
REMARK   3    DISTANCE RESTRAINTS (ANGSTROMS)
REMARK   3      BOND DISTANCE                                 0.015(0.020)
REMARK   3      ANGLE DISTANCE                                0.035(0.030)
REMARK   3      PLANAR 1-4 DISTANCE                           0.055(0.050)
REMARK   3    PLANE RESTRAINT (ANGSTROMS)                     0.021(0.020)
REMARK   3    CHIRAL-CENTER RESTRAINT (ANGSTROMS**3)          0.131(0.150)
REMARK   3    NON-BONDED CONTACT RESTRAINTS (ANGSTROMS)
REMARK   3      SINGLE TORSION CONTACT                        0.165(0.500)
REMARK   3      MULTIPLE TORSION CONTACT                      0.174(0.500)
REMARK   3      POSSIBLE HYDROGEN BOND                        0.180(0.500)
REMARK   3    CONFORMATIONAL TORSION ANGLE RESTRAINT (DEGREES)
REMARK   3      PLANAR (OMEGA)                                 4.0(3.0)
REMARK   3      STAGGERED                                     16.3(15.0)
REMARK   3      ORTHONORMAL                                   11.7(20.0)
REMARK   3    ISOTROPIC THERMAL FACTOR RESTRAINTS (ANGSTROMS**2)
REMARK   3      MAIN-CHAIN BOND                               1.38(1.000)
REMARK   3      MAIN-CHAIN ANGLE                              2.28(1.000)
REMARK   3      SIDE-CHAIN BOND                               1.97(1.000)
REMARK   3      SIDE-CHAIN ANGLE                              3.27(1.500)
REMARK   4
REMARK   4 THERE ARE TWO MOLECULES IN THE ASYMMETRIC UNIT.  THEY HAVE
REMARK   4 BEEN ASSIGNED CHAIN INDICATORS *A* AND *B*.  THEY HAVE BEEN
REMARK   4 REFINED INDEPENDENTLY WITHOUT IMPOSING NON-CRYSTALLOGRAPHIC
REMARK   4 SYMMETRY RESTRAINTS.
REMARK   5
REMARK   5 IN ADDITION TO THE METAL COORDINATION SPECIFIED ON CONECT
REMARK   5 RECORDS BELOW, THERE ARE BONDS TO OD1 AND OD2 OF ASP 10 IN
REMARK   5 A SYMMETRY-RELATED MOLECULE.  DUE TO SOME LIMITATIONS OF
REMARK   5 PROTEIN DATA BANK FORMAT, THESE BONDS CANNOT BE PRESENTED
REMARK   5 ON CONECT RECORDS.
REMARK   6
REMARK   6 CORRECTION. CORRECT CLASSIFICATION ON HEADER RECORD AND
REMARK   6  REMOVE E.C. CODE.  15-JAN-93.
SEQRES   1 A  108  SER ASP LIS ILE ILE HIS LEU TRE ASP ASP SER PHE ASP
SEQRES   2 A  108  TRE ASP VAL LEU LIS ALA ASP GLI ALA ILE LEU VAL ASP
SEQRES   3 A  108  PHE TRP ALA GLU TRP CIS GLI PRO CIS LIS MET ILE ALA
SEQRES   4 A  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TIR GLN GLI LIS
SEQRES   5 A  108  LEU TRE VAL ALA LIS LEU ASN ILE ASP GLN ASN PRO GLI
SEQRES   6 A  108  TRE ALA PRO LIS TIR GLI ILE ARG GLI ILE PRO TRE LEU
SEQRES   7 A  108  LEU LEU PHE LIS ASN GLI GLU VAL ALA ALA TRE LIS VAL
SEQRES   8 A  108  GLI ALA LEU SER LIS GLI GLN LEU LIS GLU PHE LEU ASP
SEQRES   9 A  108  ALA ASN LEU ALA
SEQRES   1 B  108  SER ASP LIS ILE ILE HIS LEU TRE ASP ASP SER PHE ASP
SEQRES   2 B  108  TRE ASP VAL LEU LYS ALA ASP GLI ALA ILE LEU VAL ASP
SEQRES   3 B  108  PHE TRP ALA GLU TRP CIS GLI PRO CIS LIS MET ILE ALA
SEQRES   4 B  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TIR GLN GLI LIS
SEQRES   5 B  108  LEU TRE VAL ALA LIS LEU ASN ILE ASP GLN ASN PRO GLY
SEQRES   6 B  108  TRE ALA PRO LIS TIR GLI ILE ARG GLI ILE PRO TRE LEU
SEQRES   7 B  108  LEU LEU PHE LIS ASN GLI GLU VAL ALA ALA TRE LIS VAL
SEQRES   8 B  108  GLI ALA LEU SER LIS GLI GLN LEU LIS GLU PHE LEU ASP
SEQRES   9 B  108  ALA ASN LEU ALA
FTNOTE   1
FTNOTE   1 RESIDUES PRO A 76 AND PRO B 76 ARE CIS PROLINES.
FTNOTE   2
FTNOTE   2 RESIDUES HIS A 6, LEU A 7, ILE A 23, ASP A 47, GLU A 48,
FTNOTE   2 LEU A 58, LEU A 80, HIS B 6, ASP B 47, LEU B 58, AND
FTNOTE   2 LEU B 80 HAVE BEEN MODELED AS TWO CONFORMERS.
FTNOTE   3
FTNOTE   3 RESIDUES 11 - 21 IN CHAIN B ARE DISORDERED.
HET     CU    109       1      COPPER ++ ION
HET     CU    109       1      COPPER ++ ION
HET    MPD    601       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    602       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    603       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    604       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    605       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    606       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    607       8      2-METHYL-2,4-PENTANEDIOL
HET    MPD    608       8      2-METHYL-2,4-PENTANEDIOL
FORMUL   3   CU    2(CU1 ++)
FORMUL   4  MPD    8(C6 H14 O2)
FORMUL   5  HOH   *140(H2 O1)
HELIX    1 A1A SER A   11  LEU A   17  1 DISORDERED IN MOLECULE B
HELIX    2 A2A CIS A   32  TIR A   49  1 BENT BY 30 DEGREES AT RES 39
HELIX    3 A3A ASN A   59  ASN A   63  1
HELIX    4 31A TRE A   66  TIR A   70  5 DISTORTED H-BONDING C-TERMINS
HELIX    5 A4A SER A   95  LEU A  107  1
HELIX    6 A1B SER B   11  LEU B   17  1 DISORDERED IN MOLECULE B
HELIX    7 A2B CIS B   32  TIR B   49  1 BENT BY 30 DEGREES AT RES 39
HELIX    8 A3B ASN B   59  ASN B   63  1
HELIX    9 31B TRE B   66  TIR B   70  5 DISTORTED H-BONDING C-TERMINS
HELIX   10 A4B SER B   95  LEU B  107  1
```

```
SHEET    1 B1A 5 LIS A   3  TRE A   8  0
SHEET    2 B1A 5 LEU A  53  ASN A  59  1  O  VAL A  55   N  ILE A   5
SHEET    3 B1A 5 GLI A  21  TRP A  28  1  N  TRP A  28   O  LEU A  58
SHEET    4 B1A 5 PRO A  76  LIS A  82 -1  O  TRE A  77   N  PHE A  27
SHEET    5 B1A 5 VAL A  86  GLI A  92 -1  N  GLI A  92   O  LIS A  82
SHEET    1 B1B 5 LIS B   3  TRE B   8  0
SHEET    2 B1B 5 LEU B  53  ASN B  59  1  O  VAL B  55   N  ILE B   5
SHEET    3 B1B 5 GLI B  21  TRP B  28  1  N  TRP B  28   O  LEU B  58
SHEET    4 B1B 5 PRO B  76  LIS B  82 -1  O  TRE B  77   N  PHE B  27
SHEET    5 B1B 5 VAL B  86  GLI B  92 -1  N  GLI B  92   O  LIS B  82
TURN     1 T1A TRE A   8  SER A  11       III (TYPE I IN MOLECULE B)
TURN     2 T2A ALA A  29  CIS A  32       I
TURN     3 T3A TIR A  49  LIS A  52       II
TURN     4 T4A GLI A  74  TRE A  77       VIB (INCLUDES CIS PRO 76)
TURN     5 T5A LIS A  82  GLU A  85       I'
TURN     6 T1B TRE B   8  SER B  11       I (TYPE III IN MOLECULE A)
TURN     7 T2B ALA B  29  CIS B  32       I
TURN     8 T3B TIR B  49  LIS B  52       II
TURN     9 T4B GLI B  74  TRE B  77       VIB (INCLUDES CIS PRO 76)
TURN    10 T5B LIS B  82  GLU B  85       I'
SSBOND   1 CIS A   32    CIS A   35
SSBOND   2 CIS B   32    CIS B   35
CRYST1   89.500   51.060   60.450  90.00 113.50  90.00 C 2          8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.011173  0.000000  0.004858        0.00000
SCALE2      0.000000  0.019585  0.000000        0.00000
SCALE3      0.000000  0.000000  0.018039        0.00000
ATOM      1  N   SER A   1      21.389  25.406  -4.628  1.00 23.22
ATOM      2  CA  SER A   1      21.628  26.691  -3.983  1.00 24.42
ATOM      3  C   SER A   1      20.937  26.944  -2.679  1.00 24.21
ATOM      4  O   SER A   1      21.072  28.079  -2.093  1.00 24.97
ATOM      5  CB  SER A   1      21.117  27.770  -5.002  1.00 28.27
ATOM      6  OG  SER A   1      22.276  27.925  -5.861  1.00 32.61
ATOM      7  N   ASP A   2      20.173  26.028  -2.163  1.00 21.39
ATOM      8  CA  ASP A   2      19.395  26.125  -0.949  1.00 21.57
ATOM      9  C   ASP A   2      20.264  26.214   0.297  1.00 20.89
ATOM     10  O   ASP A   2      19.760  26.575   1.371  1.00 21.49
ATOM     11  CB  ASP A   2      18.439  24.914  -0.856  1.00 22.14
ATOM     12  CG  ASP A   2      19.199  23.629  -0.576  1.00 23.23
ATOM     13  OD1 ASP A   2      20.107  23.371  -1.387  1.00 22.71
ATOM     14  OD2 ASP A   2      18.905  22.959   0.420  1.00 23.61

                    ... átomos da proteína foram removidos

ATOM    844  N   ALA A 108      41.357  21.341   9.676  1.00 42.93
ATOM    845  CA  ALA A 108      42.151  20.619  10.674  1.00 46.31
ATOM    846  C   ALA A 108      42.632  19.312  10.013  1.00 48.21
ATOM    847  O   ALA A 108      41.703  18.483   9.767  1.00 49.54
ATOM    848  CB  ALA A 108      41.441  20.369  11.988  1.00 46.65
ATOM    849  OXT ALA A 108      43.857  19.249   9.766  1.00 49.19
TER     850      ALA A 108

... a segunda cadeia e as moléculas de metilpentanodiol foram removidas

HETATM 1749  O   HOH   401      30.339  33.478  16.727  1.00 17.61
HETATM 1750  O   HOH   402      29.396  44.583   6.834  0.95 17.71

... 72 moléculas de água adicionais foram removidas
```

As atividades do PDB se sobrepõem com as de vários outros bancos de dados. O Cambridge Crystallographic Data Centre (CCDC) arquiva as estruturas de pequenas moléculas; oligonucleotídeos aparecem tanto no CCDC quanto no PDB. Esta informação é extremamente útil em estudos de conformações das unidades componentes de macromoléculas biológicas e nas investigações de interações macromolécula-ligante, incluindo mas não se limitando a aplicações no planejamento de fármacos. O Nucleic Acid Database (NDB) na Rutgers University, New Brunswick, New Jersey, Estados Unidos, complementa o PDB. O BioMagResBank, no Departamento de Bioquímica da University of Wisconsin, Madison, Wisconsin, Estados Unidos, arquiva estruturas de proteínas determinadas por ressonância magnética nuclear (NMR).

Os arquivos armazenam não somente os resultados de determinações de estruturas, mas também as medidas nas quais elas são baseadas. O PDB mantém

os novos dados de determinações de estruturas por difração de raios X, e o BioMagResBank, os dados obtidos por NMR.

O PDB atribui um identificador de quatro caracteres a cada uma das estruturas depositadas. O primeiro caractere é um número de 1 a 9. Não espere um significado mnemônico. Em muitos casos, diversas entradas correspondem a uma mesma proteína – determinada em diferentes estados de ligação, ou em diferentes formas cristalinas, ou resolvidas utilizando cristais de melhor qualidade ou técnicas mais apuradas de obtenção de dados. Por exemplo, existem pelo menos quatro gerações de estruturas cristalográficas da mioglobina de baleia Sperm.

É fácil recuperar uma estrutura se você conhece o seu identificador. Da *home page* do RCSB, informando como entrada um código PDB (PDB ID) e selecionando a opção *Site Search*, obtém-se uma página sumária sobre a entrada. A Figura 3.1 mostra a página sumária para a estrutura da tioredoxina, cujo identificador é 2TRX. Conexões dessa página o levam a:

PDB
PROTEIN DATA BANK

Structure Explorer - 2TRX

Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Crystallization Info

SearchLite SearchFields

Title: **Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution.**
Compound: **Thioredoxin**
Authors: **S. K. Katti, D. M. LeMaster, H. Eklund**
Exp. Method: **X-ray Diffraction**
Classification: **Electron Transport**
Source: **Escherichia coli**
Primary Citation: **Katti, S. K., LeMaster, D. M., Eklund, H.: Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution.** ***J Mol Biol*** **212** ***pp.*** **167 (1990) [Medline]**

Deposition Date: **19-Mar-1990** *Release Date:* **15-Oct-1991**

Resolution [Å]: **1.68** *R-Value:* **0.165**
Space Group: **C 2**
Unit Cell: *dim [Å]:* *a* **89.50** *b* **51.06** *c* **60.45**
angles [°]: *alpha* **90.00** *beta* **113.50** *gamma* **90.00**

Polymer Chains: **A, B** *Residues:* **216**
Atoms: **1842**
HET groups:

ID	*Name*	*Formula*
CU	**COPPER (II) ION**	CU_1
MPD	**2-METHYL-2,4-PENTANEDIOL**	$C_6H_{14}O_2$

© RCSB

Figura 3.1 Página sumária para a entrada cujo código PDB é 2TRX, a da tioredoxina de *E. coli*.*

* N. de T. Como observado em uma nota de tradução no Capítulo 1, as páginas *web* dos recursos de bioinformática estão sempre evoluindo. A que você deverá observar para o mesmo tipo de análise com a proteína cujo PDB ID é 2TRX já é um pouco diferente daquela que se vê na Figura 3.1. Entretanto, o conteúdo é, no mínimo, mais rico!

- Publicação na qual essa entrada foi descrita, via banco de dados bibliográfico como o PubMed
- Figuras da estrutura (algumas dessas podem exigir a instalação de um programa de visualização ou *plug-ins* no seu computador)
- Ao arquivo que contém as informações sobre a própria entrada
- Listas de estruturas relacionadas, de acordo com diferentes classificações de estruturas das proteínas
- Análises estereoquímicas – a distribuição de comprimentos de ligação e ângulos, e os ângulos conformacionais (de torção)
- Fontes de outras informações sobre essa entrada
- Seqüência e atribuição de estrutura secundária
- Detalhes sobre a forma do cristal e sobre os métodos utilizados para a produção dos mesmos

Tudo bem, se você conhece o identificador. Se não, como você o encontra? Da mesma forma que fizemos acima. A ferramenta simples, chamada *Site Search*, também possibilita uma pesquisa por palavras-chave. Entrando com "`coli and thioredoxin`", 262 resultados são retornados, incluindo a 2TRX e outras estruturas cristalográficas da mesma molécula ou de mutantes, mas também muitas estruturas da nuclease de estafilococos, pois, oculta nas entradas das estruturas de nucleases, está uma referência a um artigo que contém a palavra *thioredoxin* no seu título. As informações resultantes de uma pesquisa no *Site Search* lhe permitiriam verificar e facilmente escolher as estruturas a serem utilizadas na sua análise, de acordo com seus interesses nesta família de proteínas.*

O PDB também oferece navegadores mais complexos. O Macromolecular Structure Database do European Bioinformatics Institute (EBI) disponibiliza uma lista de recursos importantes para pesquisar e navegar no PDB, incluindo uma ferramenta de pesquisa chamada OCA. OCA é um navegador de banco de dados para estrutura e função de proteínas, integrando informações de inúmeros bancos de dados. Desenvolvido originalmente por J. Prilusky, OCA é mantido pelo EBI, está disponível no próprio EBI e em inúmeros sítios espelho. (O nome OCA, além de ser a palavra espanhola para ganso, tem a mesma relação com o PDB que o computador HAL de A. C. Clarke no filme 2001, Uma Odisséia no Espaço, tem com a IBM.)

Outra fonte de informações útil disponível no EBI é o banco de dados chamado Probable Quaternary Structures (PQS), de prováveis estruturas quaternárias de formas biologicamente ativas de proteínas. Em geral, a unidade assimétrica da estrutura cristalina, como depositada no PDB, contém somente parte desta unidade ativa, ou, alternativamente, cópias múltiplas da unidade ativa. Em muitos casos não fica evidente como, a partir de uma entrada específica, se obtém a sua forma ativa, e esta informação se encontra disponível no PQS.

Indicadores de qualidade da estrutura

Análises de estruturas determinadas por cristalografia por difração de raios X produzem estimativas das posições dos átomos em uma molécula e dos seus tamanhos efetivos, conhecidos como fatores de vibração térmica ou fator B (B-factor). Uma característica importante dos dados experimentais (os valores absolutos dos coeficientes de Fourier da densidade eletrônica) é que todos os

* N. de T. O resultado da análise neste parágrafo foi obtido em 26 de abril de 2007. É possível que, em uma data futura, esta mesma análise dê resultados quantitativos diferentes, ou seja, o número de entradas na saída do programa de buscas poderá ser maior ou menor. Provavelmente maior!

átomos contribuem para todas as medidas ou observações. É difícil estimar erros de posições atômicas individuais.

As determinações de estruturas cristalográficas estão à mercê do grau de ordem em diferentes partes da molécula. (Ordem é uma medida da variação entre diferentes células unitárias de um cristal.)

O grau de ordem determina a **resolução** dos dados experimentais disponíveis. A resolução é um índice de qualidade em potencial de uma estrutura determinada por difração de raios X. Ela mede a razão do número de observações pelo número de parâmetros a serem determinados. Na determinação de estruturas de pequenas moléculas orgânicas ou de minerais, essa razão geralmente é generosa: ~10. Porém, para um cristal de proteína típico, temos:

	Baixa resolução			...		**Alta**
Resolução em Å	4,0	3,5	3,0	2,5	2,0	1,5
Razão entre as observações e parâmetros	0,3	0,4	0,6	1,1	2,2	3,8

(A resolução mede a clareza com que detalhes podem ser distinguidos; assim, quanto menor o número, maior a resolução.)

Além da desordem, erros em estruturas cristalográficas refletem tanto erros nos dados quanto erros na determinação da estrutura. Uma comparação de quatro estruturas da interleucina 1β, determinadas independentemente, mostrou uma variação média de 0,84 Å na posição atômica, maior do que o erro experimental esperado.

Muitos cristalógrafos depositam seus dados experimentais junto com as estruturas resolvidas. Isso permite verificações detalhadas dos resultados. Porém, em muitos casos, os dados experimentais não estão disponíveis. Como é possível, então, que alguém avalie a qualidade de uma estrutura? Os fatores B fornecem indícios importantes; fatores B elevados em uma região sugerem que a região não foi bem determinada. Isso normalmente reflete desordem no cristal. *Softwares* especializados podem expor problemas estereoquímicos – exceções às regularidades comuns em estruturas de proteínas bem resolvidas. As entradas correspondentes às entradas do PDB em `www.cmbi.kun.nl/gv/pdbreport` descrevem análises diagnósticas, identificação de problemas e dados estranhos (*outliers*).

No entanto, embora problemas estereoquímicos sejam relativamente fáceis de *detectar*, é difícil decidir se eles realmente existem ou se correspondem, simplesmente, a características incomuns da estrutura, ou se são o resultado de erros na construção do modelo, ou ainda, o resultado inevitável de desordem do cristal. Avaliações apropriadas exigem acesso aos dados experimentais; e a correção de erros reais pode necessitar da atenção de um cristalógrafo experiente. A conclusão parece não fugir de que fatores de estrutura deveriam ser arquivados e estar disponíveis.

Ressonância magnética nuclear (NMR)

A NMR é a segunda técnica mais utilizada para a determinação de estruturas de macromoléculas. Ela geralmente produz estruturas com topologias corretas, mas não tão precisas quanto uma boa estrutura determinada por difração de raios X e, por isso, são menos importantes no estudo de detalhes refinados da estrutura. Os cristalógrafos descrevem uma única estrutura, ou somente um pequeno número de estruturas. Espectroscopistas de NMR em geral produzem uma família de aproximadamente 10 a 20 estruturas relacionadas ou até mais, calculadas a partir dos

mesmos dados experimentais. A comparação sobre um conjunto desses indica precisão; regiões nas quais a variação local na estrutura é pequena são bem definidas pelos dados. Este é um equivalente aproximado do fator B dos cristalógrafos.

Recursos na *web*: Estruturas de proteínas e ácidos nucléicos

***Home page* do Protein Data Bank:**
`http://www.rcsb.org/`

***Home page* do EBI Macromolecular Structure Database:**
`http://msd.ebi.ac.uk/`

***Home page* do BioMagResBank:**
`http://www.bmrb.wisc.edu/`

Pesquisando no Protein Data Bank:

***Home page* do SCOP (Structural Classification of Proteins):**
`http://scop.mrc-lmb.cam.ac.uk/scop/`

Lista de navegadores:
`http://pdb-browsers.ebi.ac.uk/browse_it.shtml`

OCA: `http://oca.ebi.ac.uk/oca-bin/ocamain`

Bancos de dados de estruturas quaternárias de proteínas:
`http://pqs.ebi.ac.uk/`

Informações sobre a qualidade de estruturas:
`http://www.cmbi.kun.nl/gv/pdbreport/`

Classificação de estrutura de proteínas

Muitos sítios na *web* oferecem classificações hierárquicas de todo o Protein Data Bank, de acordo com os padrões de enovelamento das proteínas (ver Capítulo 5):

- SCOP: Classificação estrutural de proteínas
- CATH: Classe/Arquitetura/Topologia/Superfamília homóloga?
- DALI: baseado na extração de estruturas similares a partir de matrizes de distância
- CE: um banco de dados de alinhamentos de estruturas

Esses sítios são pontos de entradas comuns úteis para dados estruturais de proteínas. Por exemplo, SCOP oferece recursos para pesquisa por palavras-chave para identificar estruturas, navegar acima e abaixo na hierarquia, gerar figuras, acessar registros de anotação nas entradas PDB, e ainda possui conexões para bancos de dados relacionados.

Bancos de dados especializados ou "boutique"

Muitas pessoas ou grupos selecionam, anotam e recombinam dados direcionados a assuntos específicos, e incluem conexões que permitem um acesso eficiente a informações sobre assuntos de interesse.

Por exemplo, o Protein Kinase Resource é uma compilação especializada em proteínas cinases, a qual inclui seqüências, estruturas, informações funcionais, procedimentos laboratoriais, lista de cientistas interessados, ferramentas para análise, quadro de avisos e conexões.

O banco de dados da protease do HIV arquiva as estruturas das proteinases do vírus da imunodeficiência humana tipo 1, do vírus da imunodeficiência humana tipo 2, e as do vírus da imunodeficiência Simian e seus complexos; e ainda oferece ferramentas para suas análises e conexões para outros sítios com informações relacionadas à AIDS. Esse banco de dados contém algumas estruturas cristalinas não depositadas no PDB.

No campo da imunologia:

- IMGT, o banco de dados internacional ImMunoGeneTics, é um banco de dados integrado de alta qualidade e especializado em imunoglobulinas, receptores de células-T e moléculas do complexo principal de histocompatibilidade (*Major Histocompatibility Complex*, MHC) de todas as espécies de vertebrados. O servidor IMGT oferece um acesso comum a todos os dados imunogenéticos. Atualmente ele inclui dois bancos de dados: o IMGT/LIGM-DB, um banco de dados abrangente de seqüências de genes de imunoglobulinas e de receptores de células-T de humanos e outros vertebrados, com tradução das seqüências completamente anotadas; e o IMGT/HLA-DB, um banco de dados de MHC humanos referidos como HLA (*Human Leucocyte Antigens*).
- KABAT – Banco de dados de seqüências de proteínas de interesse imunológico – North-Western University (Estados Unidos).
- MHCPEP – Um banco de dados de peptídeos que se ligam ao complexo principal de histocompatibilidade – WEHI (Melbourne, Austrália).

Recursos na *web*: Bancos de dados de famílias de proteínas específicas

Proteínas cínases:
`http://www.sdsc.edu/kinases/`

Proteases do HIV:
`http://mcl1.ncifcrf.gov/hivdb/index.html`
`http://xpdb.nist.gov/hivsdb/hivsdb.html`

Imunologia:
IGMT: `http://imgt.cines.fr`
KABAT: `http://immuno.bme.nwu.edu/`
MHCPEP: `http://wehih.whi.edu.au/mhcpep/`

Bancos de dados de padrões de expressão e de proteômica

Lembre-se do dogma central: DNA produz RNA que produz proteína. Bancos de dados genômicos contêm seqüências de DNA. Bancos de dados de expressão armazenam as medições dos níveis de *mRNA*, geralmente por meio das ESTs (etiquetas de seqüências expressas – seqüências terminais curtas de cDNA sintetizadas a partir de mRNA) descrevendo padrões de transcrição de genes. Bancos de dados de proteômica armazenam medidas realizadas com *proteínas*, descrevendo padrões de tradução de genes.

Comparações de padrões de expressão fornecem indícios de (1) função e mecanismo de ação de produtos gênicos, (2) como os organismos coordenam o controle interno sobre os seus processos metabólicos em diferentes condições – por exemplo, a levedura sob condições aeróbias ou anaeróbias, (3) variações na mobilização de genes durante os diferentes estágios do ciclo celular, ou do desenvolvimento de um organismo, (4) mecanismos de resistência a antibió-

ticos em bactérias, e a conseqüente proposta de alvos para o desenvolvimento de fármacos, (5) resposta ao desafio por um parasita, (6) resposta a diferentes tipos e dosagens de medicamentos, para guiar uma terapia efetiva.

Existem muitos bancos de dados de ESTs. Em sua maioria, as entradas contêm campos que indicam os tecidos de origem e/ou a localização subcelular, o estágio do desenvolvimento, as condições de crescimento e a quantificação dos níveis de expressão. Dentro do GenBank, o dbEST contém atualmente quase 43 milhões de entradas, de 1.200 espécies*, lideradas por:

Espécies com o maior número de entradas no dbEST*

Espécies	Número de entradas
Homo sapiens (humanos)	7.974.440
Mus musculus + *domesticus* (camundongo)	4.840.615
Danio rerio (peixe-zebra)	1.350.085
Bos taurus (gado)	1.315.093
Arabidopsis thaliana (agrião thale)	1.276.131
Xenopus tropicalis (rã)	1.256.244
Oryza sativa (arroz)	1.211.154
Zea mays (milho)	1.161.241
Triticum aestivum (trigo)	1.050.131
Rattus sp. (rato)	871.163
Ciona intestinalis (cordados primitivos)	686.396
Xenopus laevis (*rã africana*)	677.784
Sus scrofa (javali)	641.896
Gallus gallus (galinha)	599.330
Drosophila melanogaster (mosca-das-frutas)	532.557
Hordeum vulgare + subsp. *vulgare* (cevada)	437.728
Glycine max (soja)	371.817
Caenorhabditis elegans (nematódeo)	346.064

Alguns dos conjuntos de dados de EST são especializados em tecidos específicos (por exemplo, músculos, dentes) ou em espécies. Em muitos casos, há esforços para relacionar padrões de expressão a outras informações conhecidas sobre o organismo. Por exemplo, o Gene Expression Information Resource Project for Mouse Development do Jackson Laboratory, Estados Unidos, coordena dados sobre expressão de genes e anatomia do desenvolvimento de camundongos.

Muitos bancos de dados oferecem conexões entre ESTs de diferentes espécies, por exemplo, ligando homólogos humanos e de camundongos, ou relações entre genes de doenças humanas e proteínas de levedura. Outros conjuntos de dados EST são especializados em um tipo de proteína, por exemplo, citocinas. Um grande esforço está centrado no câncer: integrando informações sobre mutantes, rearranjos cromossômicos e mudanças nos padrões de expressão de genes para identificar alterações genéticas durante a formação e progressão do tumor.

É claro que, embora exista uma estreita relação entre padrões de transcrição e padrões de tradução, quantificações diretas do conteúdo de proteínas de células e tecidos – proteômica – oferecem informações adicionais valiosas. Devido às

* N. de T. Os números deste parágrafo e da tabela foram atualizados com base nos dados do dbEST em 13 de abril de 2007.

diferentes taxas de tradução de diferentes mRNAs, a quantificação direta de proteínas fornece uma descrição mais acurada de padrões de expressão de genes do que a quantificação da transcrição. Modificações pós-traducionais podem ser detectadas *somente* examinando as proteínas.

A análise proteômica envolve a separação, a identificação e a determinação das quantidades de proteínas em uma amostra (ver Capítulo 6). Bancos de dados de proteômica armazenam imagens de géis e sua interpretação em termos de padrões de proteínas. Para cada proteína, uma entrada tipicamente registra (ver *Weblema* 3.21):

- identificação da proteína
- quantidade relativa
- função
- mecanismo de ação
- padrão de expressão
- localização subcelular
- proteínas relacionadas
- modificações pós-traducionais
- interações com outras proteínas
- conexões com outros bancos de dados

A bioinformática está contribuindo para o desenvolvimento desses bancos de dados, e também para o desenvolvimento de algoritmos de comparação e análise de padrões que eles contenham.

Bancos de dados de vias metabólicas

A Kyoto Encyclopedia of Genes and Genomes (KEGG) arquiva genomas individuais, produtos de genes e suas funções, mas o seu diferencial principal está na sua integração de informações bioquímicas e genéticas. A KEGG se concentra nas interações: associações de moléculas e redes metabólicas e reguladoras. Ela está sendo desenvolvida sob a direção de M. Kanehisa.

A KEGG organiza cinco tipos de dados em um sistema que compreende:

1. Catálogos de compostos químicos em células vivas
2. Catálogos de genes
3. Mapas de genomas
4. Mapas de vias
5. Tabelas de ortólogos

Os catálogos de compostos químicos e genes – itens 1 e 2 – contêm informações sobre moléculas ou seqüências específicas. O item 3, mapas de genomas, integra os próprios genes de acordo com as suas localizações nos cromossomos. Em alguns casos, sabendo-se que um gene aparece em um óperon, é possível obter indícios sobre a sua função.

No item 4, os mapas de vias descrevem redes em potencial de atividades moleculares, tanto metabólicas quanto reguladoras. Uma via metabólica na KEGG é uma idealização correspondendo a um grande número de possíveis cascatas metabólicas. Ela pode gerar uma via metabólica real de um organismo particular, alinhando as proteínas daquele organismo com enzimas dentro das vias de referência.

Uma enzima em um organismo é referida na KEGG em suas tabelas de ortólogos, item 5, que liga a enzima a outras relacionadas, presentes em outros

organismos. Isso permite a análise das relações entre as vias metabólicas de diferentes organismos.

O poder da KEGG provém de uma rede muito densa de conexões entre essas categorias de informações, além de conexões adicionais para muitos outros bancos de dados com os quais o sistema mantém acesso. Dois exemplos de tipos de questões que podem ser tratados pela KEGG são:

- Sugeriu-se que vias metabólicas simples evoluem para vias mais complexas através da duplicação de genes e subseqüente divergência. A procura no catálogo de vias por conjuntos de enzimas que compartilham um padrão de enovelamento revelará grupos de parálogas.
- A KEGG pode utilizar um conjunto de enzimas de um determinado organismo e verificar se elas podem ser integradas em vias metabólicas conhecidas. Uma brecha em uma via sugere que uma enzima foi perdida ou que surgiu uma via alternativa não esperada.

Bancos de dados bibliográficos

O MEDLINE (baseado na US National Library of Medicine) integra a literatura biomédica, incluindo muitos artigos lidando com assuntos sobre biologia molecular, mas de conteúdo não demasiadamente clínico. Ele faz parte do PubMed, um banco de dados bibliográfico que oferece resumos de artigos científicos, integrado com outras ferramentas de recuperação de informação do NCBI (`http://www.ncbi.nlm.nih.gov/PubMed/`).

Uma característica bastante eficiente do PubMed é a opção de buscar por artigos relacionados. Esta é uma maneira muito rápida de "ser introduzido" à literatura sobre um tópico. Combinadas com o uso de um instrumento de busca geral de sítios na *web* que não correspondam a artigos publicados em periódicos, informações razoavelmente abrangentes sobre a maioria dos assuntos se tornam facilmente disponíveis. Aqui vai uma dica: se você está tentando começar a aprender sobre um assunto com o qual não tem familiaridade, tente adicionar a palavra-chave *tutorial* em sua pesquisa com uma ferramenta de busca geral, ou a palavra-chave *review* (revisão) em sua busca no PubMed.*

Atualmente, quase todos os periódicos científicos disponibilizam suas listas de conteúdos e, em muitos casos, suas edições completas, em sítios na *web*. O US National Institutes of Health estabeleceu uma biblioteca de artigos científicos centralizada e baseada na *web*, chamada PubMed Central (`http://www.pubmedcentral.nih.gov/`). Em colaboração com periódicos científicos, o NCBI está organizando a distribuição eletrônica dos textos completos de artigos publicados.

Uma nova organização, a Public Library of Science (PLoS), tem o objetivo de tornar a literatura científica (incluindo a médica) pública e de livre acesso. Como é uma organização sem fins lucrativos, a PLoS tem recebido ajuda de fundações para os seus esforços para distribuir a literatura publicada por outros e para começar suas próprias publicações, o que permitirá uma exploração de diferentes relacionamentos – incluindo mas não se limitando aos econômicos – entre autores, editores e leitores.

Avaliações sobre bancos de dados e servidores de biologia molecular

É difícil explorar na *web* qualquer tópico de biologia molecular sem se deparar rapidamente com uma lista de bancos de dados e servidores de biologia mole-

* N. de T. Atualmente, as pesquisas no PubMed incluem, automaticamente, como parte da lista de resultados, uma conexão apenas para artigos de revisão sobre o assunto que está sendo pesquisado.

cular. Essas listas de recursos na *web* de biologia molecular são muito comuns. Eles contêm, até certo ponto, a mesma informação, mas as suas interfaces variam consideravelmente. O único problema é que, a menos que esses recursos sejam organizados e administrados, eles tendem a se degenerar em listas de conexões perdidas. (Um rascunho desta seção continha uma referência a um sítio da *web* com uma relação apreciável de sítios. Retornando a ele dois meses depois, o nome do sítio havia se modificado, e mais da metade dos sítios listados haviam desaparecido.)

Este livro não contém uma lista longa e comentada de sítios relevantes e recomendados pelas seguintes razões: (1) Você não quer uma lista longa, você precisa de uma curta. (2) A *web* é muito instável para que tal lista de sítios permaneça útil por bastante tempo. *É muito mais eficiente utilizar um instrumento de busca geral para encontrar o que você quer no momento que você necessita.* A cada ano, a edição de janeiro da revista Nucleic Acids Research publica uma coleção de artigos sobre bancos de dados de biologia molecular. Esta é uma referência inestimável.

Além disso, o conteúdo dos bancos de dados está se expandindo continuamente. Se você testar as pesquisas descritas nos exemplos deste capítulo, você obterá mais resultados do que os impressos aqui. (Na verdade, eu não hesitei em utilizar resultados antigos, pois eles apresentam uma variedade maior do que os resultados mais recentes e parecem mais informativos. O problema de suprimir a redundância crescente em respostas a consultas na *web* é um desafio para as pesquisas na área de recuperação de informação.)

Meu conselho é: gaste algum tempo navegando; não levará muito tempo para que você encontre um sítio que pareça razoavelmente estável e que tenha um estilo parecido com seus métodos de trabalho. Alternativamente, aqui está um sítio que é completo e que apresenta sinais de comprometimento em se manter sempre atualizado: `http://www.expasy.org/alinks.html`. Este é um sítio apropriado para se iniciar uma sessão de navegação em bioinformática de análise de seqüências de proteínas e ácidos nucléicos.

Acessos a bancos de dados

Os bancos de dados de seqüências de ácidos nucléicos e de proteínas mantêm recursos para uma ampla variedade de mecanismos de recuperação de informação e realização de análises. Categorias de possibilidades de análises incluem:

1. **Recuperação de seqüências em bancos de dados** Seqüências podem ser "convocadas" com base nas características das suas anotações ou por meio de padrões encontrados nas próprias seqüências.
2. **Comparação de seqüências** Isto não é um recurso, é uma indústria pesada! Ela foi introduzida no Capítulo 1 e será discutida no Capítulo 4. A comparação de seqüências inclui pesquisas importantes sobre homologias.
3. **Tradução de seqüências de DNA em seqüências de proteínas**
4. **Tipos simples de análise e predição de estruturas** Por exemplo, métodos estatísticos para predição de estrutura secundária de proteínas a partir somente de seqüências, incluindo perfis hidrofóbicos – a partir dos quais as proteínas transmembranas geralmente podem ser identificadas (ver página 213).
5. **Reconhecimento de padrões** É possível pesquisar por todas as seqüências contendo um padrão ou combinação de padrões, expressos como probabi-

lidades de se encontrar certos grupos de resíduos em posições consecutivas. Em seqüências de DNA, estes podem ser sítios de reconhecimento para enzimas, como aquelas responsáveis por recomporem genes interrompidos. Em proteínas, padrões curtos e localizados, algumas vezes, identificam moléculas que compartilham uma mesma função, ainda que não exista uma relação óbvia entre suas seqüências. PROSITE é uma coleção desses padrões de "assinatura" de proteínas.

6. **Gráficos moleculares** são necessários para fornecer descrições inteligíveis de sistemas muito complicados. Aplicações típicas de gráficos moleculares incluem:
   - O mapeamento de resíduos, que se acredita estarem envolvidos na função, na estrutura tridimensional da proteína. Isso em geral mostrará o sítio ativo da proteína mais claramente.
   - A classificação e a comparação de padrões de enovelamento de proteínas.
   - A análise das mudanças entre estruturas estreitamente relacionadas ou entre dois estados conformacionais de uma única molécula.
   - O estudo da interação de uma pequena molécula com uma proteína, com o objetivo de tentar atribuir sua função ou de desenvolver fármacos.
   - O ajuste interativo de um modelo à imagem indefinida de uma molécula, obtida inicialmente pela cristalografia por difração de raios X, para a solução da estrutura de proteínas.
   - A modelagem e o projeto de novas estruturas.

Acesso a bancos de dados de biologia molecular

Como adquirir habilidades para trabalhar na *web*

Seria difícil aprender a andar de bicicleta lendo um livro que descrevesse o conjunto de movimentos necessários, muito menos lendo um sobre a teoria do giroscópio. De maneira similar, o lugar certo para adquirir habilidades na *web* é em um terminal de computador, utilizando um navegador. Isso é verdade, mas há sempre um certo período de dificuldade e ansiedade. Aqui o objetivo é somente oferecer alguma assistência temporária para que você consiga começar. Então, você pode seguir em frente!

Esta seção contém introduções para alguns dos principais bancos de dados e sistemas de recuperação de informação da área de biologia molecular. Em cada caso, mostraremos aplicações e pesquisas relativamente simples. Quando apropriado, características exclusivas de cada sistema serão enfatizadas.

ENTREZ

O NCBI, um componente da United States National Library of Medicine, mantém bancos de dados e meios de acesso a esses dados. O sistema ENTREZ oferece acesso por meio das seguintes divisões de bancos de dados:

- Proteína
- Peptídeo
- Nucleotídeo
- Gene
- Estrutura
- Genoma

- Popset – informações sobre populações
- OMIM – Online Mendelian Inheritance in Man

Um dos pontos fortes do sistema do NCBI são as conexões entre os vários bancos de dados. O ponto inicial para a recuperação de seqüências e estruturas é chamado ENTREZ: `http://www.ncbi.nlm.nih.gov/Entrez/`.

Vamos escolher uma molécula – elastase de neutrófilos humana (do inglês, *human neutrophil elastase*) – e pesquisar por entradas relevantes em diferentes seções do ENTREZ.

Pesquisa no banco de dados de proteínas utilizando o ENTREZ

Vá para `http://www.ncbi.nlm.nih.gov/Entrez/`. Selecione `Protein: Sequence Database`, e entre com os termos de pesquisa: HUMAN ELASTASE e clique em GO.

O Quadro mostra quinze respostas retornadas pelo programa.* (Em um navegador, você também encontrará conexões para entradas em bancos de dados de seqüências.) O primeiro resultado é ELASTASE 1 PRECURSOR [HOMO SAPIENS]; outras respostas incluem elastases de outras espécies, inibidores de humanos e de sanguessugas, e tirosil-tRNA sintetase. (Por que uma proteína de sanguessuga e uma tRNA sintetase apareceriam em uma pesquisa por elastase humana? Ver *Weblema* 3.9.) Depois veremos como adequar a pesquisa para eliminar essas respostas irrelevantes.

Resultados do ENTREZ para uma busca por *elastase humana* no banco de dados PROTEIN

1. elastase 1 precursor [Homo sapiens]
   gi—4731318—gb—AAD28441.1—AF120493_1[4731318]
2. ALPHA-1-ANTITRYPSIN PRECURSOR (ALPHA-1 PROTEASE INHIBITOR) (ALPHA-1-ANTIPROTEINASE)
   gi—1703025—sp—P01009—A1AT_HUMAN[1703025]
3. elastase [Mus musculus]
   gi—7657060—ref—NP_056594.1—[7657060]
4. proteinase 3 [Mus musculus]
   gi—6755184—ref—NP_035308.1—[6755184]
5. ANTIMICROBIAL PEPTIDE ENAP-2
   gi—7674025—sp—P56928—ENA2_HORSE[7674025]
6. AMBP PROTEIN PRECURSOR [CONTAINS: ALPHA-1-MICROGLOBULIN (PROTEIN HC) (COMPLEX-FORMING GLYCOPROTEIN HETEROGENEOUS IN CHARGE); INTER-ALPHA-TRYPSIN INHIBITOR LIGHT CHAIN (ITI-LC) (BIKUNIN) (HI-30)]
   gi—122801—sp—P02760—AMBP_HUMAN[122801]

→

* N. de T. Como mencionado no Capítulo 1 pelo tradutor e advertido pelo autor, na página 160, a lista de resultados no Quadro é diferente quando a mesma busca foi realizada em 27 de abril de 2007. Provavelmente, ela será diferente quando você repetir este exercício! E lembre-se, diferente não é necessariamente errado!

7. ELAFIN PRECURSOR (ELASTASE-SPECIFIC INHIBITOR) (ESI) (SKIN-DERIVED ANTILEUKOPROTEINASE) (SKALP)
gi—119262—sp—P19957—ELAF_HUMAN[119262]
8. ANTILEUKOPROTEINASE
gi—113637—sp—P22298—ALK1_PIG[113637]
9. ANTILEUKOPROTEINASE 1 PRECURSOR (ALP) (HUSI-1) (SEMINAL PROTEINASE INHIBITOR) (SECRETORY LEUKOCYTE PROTEASE INHIBITOR) (BLPI) (MUCUS PROTEINASE INHIBITOR) (MPI)
gi—113636—sp—P03973—ALK1_HUMAN[113636]
10. ALPHA-2-MACROGLOBULIN PRECURSOR (ALPHA-2-M)
gi—112911—sp—P01023—A2MG_HUMAN[112911]
11. tyrosyl-tRNA synthetase [Homo sapiens]
gi—4507947—ref—NP_003671.1—[4507947]
12. pancreatic elastase IIB [Homo sapiens]
gi—7705648—ref—NP_056933.1—[7705648]
13. protease inhibitor 3, skin-derived (SKALP) [Homo sapiens]
gi—4505787—ref—NP_002629.1—[4505787]
14. pancreatic elastase I (allele HEL1-36)—human (fragment)
gi—7513237—pir—S70441[7513237]
15. guamerin—Korean leech

As respostas têm o seguinte formato: em cada caso, a primeira linha mostra o nome e os sinônimos para a molécula, e a espécie de origem. Note que letras gregas são soletradas. A última linha mostra referências aos bancos de dados utilizados como fonte dessas informações: gi = identificador GenInfo (ver página 45), gb = número de acesso ao GenBank, sp = número de acesso ao Swiss-Prot, pir = número de acesso ao Protein Information Resource, ref = número de acesso ao projeto Reference Sequence do NCBI. As entradas retornadas incluem elastases de humanos e de outras espécies, assim como inibidores de elastases.

Abrindo a entrada correspondente ao primeiro resultado retornado, obtém-se o arquivo mostrado no próximo Quadro. As primeiras linhas são, principalmente, informações relacionadas ao gerenciamento interno do banco de dados – número de acesso, nome da molécula, data de depósito, etc. Então, aparecem informações descritivas, como a origem dos dados, neste caso, humana, com a classificação taxonômica completa, créditos aos cientistas que depositaram a entrada e referências da literatura. Finalmente, as informações científicas específicas: a localização do gene e seu produto (CDS = seqüência codificadora), e a própria seqüência (ver Exercício 3.2).

Pesquisas no banco de dados de nucleotídeos utilizando o ENTREZ

Vamos novamente buscar por HUMAN ELASTASE, mas desta vez no banco de dados de Nucleotide. Vamos tentar adequar a pesquisa para eliminar as respostas que se referem a inibidores da elastase.

1. Selecione NUCLEOTIDE no sítio do ENTREZ.
2. Digite HOMO SAPIENS [Organism] AND ELASTASE 1 [Title] NOT INHIBITOR [Text Word] no campo de pesquisa.
3. Clique em GO.

Primeiro resultado da pesquisa por elastase humana no banco de dados PROTEIN utilizando o ENTREZ

```
LOCUS       AF120493_1     258 aa                      PRI       03-AUG-2000
DEFINITION  elastase 1 precursor [Homo sapiens].
ACCESSION   AAD28441
PID         g4731318
VERSION     AAD28441.1  GI:4731318
DBSOURCE    locus AF120493 accession AF120493.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 258)
  AUTHORS   Talas,U., Dunlop,J., Khalaf,S., Leigh,I.M. and Kelsell,D.P.
  TITLE     Human elastase 1: evidence for expression in the skin and the
            identification of a frequent frameshift polymorphism
  JOURNAL   J. Invest. Dermatol. 114 (1), 165-170 (2000)
  MEDLINE   20087075
   PUBMED   10620133
REFERENCE   2  (residues 1 to 258)
  AUTHORS   Talas,U., Dunlop,J., Leigh,I.M. and Kelsell,D.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-JAN-1999) Centre for Cutaneous Research, Queen Mary
            and Westfield College, 2 Newark Street, London E1 2AT, UK
COMMENT     Method: conceptual translation supplied by author.
FEATURES             Location/Qualifiers
     source          1..258
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="12"
                     /map="12q13"
                     /cell_type="keratinocyte"
     Protein         1..258
                     /product="elastase 1 precursor"
     CDS             1..258
                     /gene="ELA1"
                     /coded_by="AF120493.1:42..818"
ORIGIN
        1 mlvlyghstq dlpetnarvv ggteagrnsw psqislqyrs ggsryhtcgg tlirqnwvmt
       61 aahcvdyqkt frvvagdhnl sqndgteqyv svqkivvhpy wnsdnvaagy diallrlaqs
      121 vtlnsyvqlg vlpqegaila nnspcyitgw gktktngqla qtlqqaylps vdyaicssss
      181 ywgstvkntm vcaggdgvrs gcqgdsggpl hclvngkysl hgvtsfvssr gcnvsrkptv
      241 ftqvsayisw innviasn
//
```

A pesquisa retorna nove resultados, incluindo vários clones individuais. O primeiro resultado (ver Quadro a seguir) é: HOMO SAPIENS ELASTASE 1 PRECURSOR (ELA1) mRNA, COMPLETE CDS. O termo COMPLETE CDS significa uma seqüência codificadora completa. Se utilizássemos apenas ELASTASE em vez de ELASTASE 1, obteríamos mais de 800 resultados.*

Compare este arquivo com o resultado da pesquisa no banco de dados PROTEIN (ver Exercício 3.5).

* N. de T. Ver Nota de Tradutor da p. 162.

Primeiro resultado da pesquisa por elastase humana no banco de dados NUCLEOTIDE utilizando o ENTREZ

```
LOCUS       AF120493      952 bp    mRNA            PRI       03-AUG-2000
DEFINITION  Homo sapiens elastase 1 precursor (ELA1) mRNA, complete cds.
ACCESSION   AF120493
VERSION     AF120493.1  GI:4731317
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 952)
  AUTHORS   Talas,U., Dunlop,J., Khalaf,S., Leigh,I.M. and Kelsell,D.P.
  TITLE     Human elastase 1: evidence for expression in the skin and the
            identification of a frequent frameshift polymorphism
  JOURNAL   J. Invest. Dermatol. 114 (1), 165-170 (2000)
  MEDLINE   20087075
   PUBMED   10620133
REFERENCE   2  (bases 1 to 952)
  AUTHORS   Talas,U., Dunlop,J., Leigh,I.M. and Kelsell,D.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-JAN-1999) Centre for Cutaneous Research, Queen Mary
            and Westfield College, 2 Newark Street, London E1 2AT, UK
FEATURES             Location/Qualifiers
     source          1..952
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="12"
                     /map="12q13"
                     /cell_type="keratinocyte"
     gene            1..952
                     /gene="ELA1"
     CDS             42..818
                     /gene="ELA1"
                     /codon_start=1
                     /product="elastase 1 precursor"
                     /protein_id="AAD28441.1"
                     /db_xref="GI:4731318"
                     /translation="MLVLYGHSTQDLPETNARVVGGTEAGRNSWPSQISLQYRSGGSR
                     YHTCGGTLIRQNWVMTAAHCVDYQKTFRVVAGDHNLSQNDGTEQYVSVQKIVVHPYWN
                     SDNVAAGYDIALLRLAQSVTLNSYVQLGVLPQEGAILANNSPCYITGWGKTKTNGQLA
                     QTLQQAYLPSVDYAICSSSSYWGSTVKNTMVCAGGDGVRSGCQGDSGGPLHCLVNGKY
                     SLHGVTSFVSSRGCNVSRKPTVFTQVSAYISWINNVIASN"
BASE COUNT      226 a    261 c    250 g    215 t
ORIGIN
        1 ttggtccaag caagaaggca gtggtctact ccatcggcaa catgctggtc ctttatggac
       61 acagcaccca ggaccttccg gaaaccaatg cccgcgtagt cggagggact gaggccggga
      121 ggaattcctg gccctctcag atttccctcc agtaccggtc tggaggttcc cggtatcaca
      181 cctgtggagg gacccttatc agacagaact gggtgatgac agctgctcac tgcgtggatt
      241 accagaagac tttccgcgtg gtggctggag accataacct gagccagaat gatggcactg
      301 agcagtacgt gagtgtgcag aagatcgtgg tgcatccata ctggaacagc gataacgtgg
      361 ctgccggcta tgacatcgcc ctgctgcgcc tggcccagag cgttaccctc aatagctatg
      421 tccagctggg tgttctgccc caggagggag ccatcctggc taacaacagt ccctgctaca
      481 tcacaggctg gggcaagacc aagaccaatg ggcagctggc ccagaccctg cagcaggctt
      541 acctgccctc tgtggactat gccatctgct ccagctcctc ctactggggc tccactgtga
      601 agaacaccat ggtgtgtgct ggtggagatg gagttcgctc tggatgccag ggtgactctg
      661 ggggcccccc ccattgcttg gtgaatggca agtattctct ccatggagtg accagctttg
      721 tgtccagccg gggctgtaat gtctccagga agcctacagt cttcacccag gtctctgctt
      781 acatctcctg gataaataat gtcatcgcct ccaactgaac attttcctga gtccaacgac
      841 cttcccaaaa tggttcttag atctgcaata ggacttgcga tcaaaaagta aaacacattc
      901 tgaaagacta ttgagccatt gatagaaaag caaataaaac tagatataca tt
//
```

Pesquisas no banco de dados de genoma utilizando o ENTREZ

Uma pesquisa por HUMAN ELASTASE retorna:

1. NC_000967 CAENORHABDITIS ELEGANS CHROMOSOME III[64] LCL—WORM_CHR_III
2. NC_001099 HOMO SAPIENS CHROMOSOME 19[19] REF—NC_001099—HSAP-19
3. NC_001065 HOMO SAPIENS CHROMOSOME 14[14] REF—NC_001065—HSAP-14
4. NC_001044 HOMO SAPIENS CHROMOSOME 11[11] REF—NC_001044—HSAP-11
5. NC_001008 HOMO SAPIENS CHROMOSOME 6[6] REF—NC_001008—HSAP-6

Por que uma proteína de *C. elegans* apareceria em uma pesquisa por elastase humana? A entrada NC_000967 é o cromossomo III completo de *C. elegans*. Comentários em um dos genes detectados incluem:

```
Gene="T07A5.1" /note="weak similarity with elastase (PIR Accession number A406659)"
```

Muitos outros genes de *C. elegans* estão anotados como similares a proteínas humanas. Entretanto, apesar de *C. elegans* conter uma elastase, esta *não* é assinalada como similar à elastase humana, embora seja uma homóloga.

Pesquisas em bancos de dados de estruturas utilizando o ENTREZ

A estrutura tridimensional da elastase humana é conhecida? Selecione o banco de dados STRUCTURE dentre as opções do campo *Search* e reexecute a pesquisa com a seguinte expressão no campo de consulta: HOMO SAPIENS [Organism] AND ELASTASE [Title] NOT INHIBITOR [Text Word]. Note que desta vez estamos utilizando ELASTASE e não ELASTASE 1. O programa retorna pelo menos cinco respostas:

1JK3	CRYSTAL STRUCTURE OF HUMAN MMP-12 (MACROPHAGE ELASTASE) AT TRUE ATOMIC RESOLUTION
1HAZ	SNAPSHOTS OF SERINE PROTEASE CATALYSIS: (C) ACYL-ENZYME INTERMEDIATE BETWEEN PORCINE PANCREATIC ELASTASE AND HUMAN BETA-CASOMORPHIN-7 JUMPED TO PH 9 FOR 1 MINUTE
1HAX	SNAPSHOTS OF SERINE PROTEASE CATALYSIS: (A) ACYL-ENZYME INTERMEDIATE BETWEEN PORCINE PANCREATIC ELASTASE AND HUMAN BETA-CASOMORPHIN-7 AT PH 5
1BOF	CRYSTAL STRUCTURE OF HUMAN NEUTROPHIL ELASTASE WITH MDL 101, 146
1QIX	PORCINE PANCREATIC ELASTASE COMPLEXED WITH HUMAN BETA-CASOMORPHIN-7

As designações 1HAX, 1JK3, 1HAZ, 1BOF e 1QIX são códigos de entrada do Protein Data Bank.

OOPS! – podemos não ter percebido, mas perdemos muitas entradas importantes. Existem muitas estruturas de elastases resolvidas em complexos com inibidores, os quais havíamos solicitado que o sistema rejeitasse. Removendo NOT INHIBITORS e reexecutando a consulta, a pesquisa retorna muito mais estruturas. Experimente!

Pesquisas no banco de dados bibliográficos PubMed

Talvez este seja o momento de procurar pelo que as pessoas têm para dizer sobre a nossa molécula. É claro que a literatura sobre a elastase é vasta. Uma pesquisa por HUMAN ELASTASE no PubMed retorna mais de 7.500 entradas. Para suprimir resultados supérfluos, vamos tentar encontrar citações de artigos que descrevam a atuação da elastase em doenças. Uma pesquisa por HUMAN ELASTASE DISEASE retorna mais de 1.600 resultados. Que tal procurar por **mutantes** específicos da elastase relacionados com doenças humanas? Uma pesquisa por HUMAN ELASTASE DISEASE MUTATION retorna mais de 40 artigos, em ordem cronológica inversa. A seguir, encontra-se uma lista de 10 desses artigos:

Hermans MH, Touw IP. Significance of neutrophil elastase mutations versus G-CSF receptor mutations for leukemic progression of congenital neutropenia. Blood. 2001 Apr 1;97(7):2185–6. No abstract available.

Li FQ , Horwitz M. Characterization of mutant neutrophil elastase in severe congenital neutropenia. J Biol Chem. 2001 Apr 27;276(17):14230–41.

Ye S. Polymorphism in matrix metalloproteinase gene promoters: implication in regulation of gene expression and susceptibility of various diseases. Matrix Biol. 2000 Dec;19(7):623–9. Review.

Dale DC, Person RE, Bolyard AA, Aprikyan AG, Bos C, Bonilla MA, Boxer LA, Kannourakis G, Zeidler C, Welte K, Benson KF, Horwitz M. Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. Blood. 2000 Oct 1;96(7):2317–22.

McGettrick AJ, Knott V, Willis A, Handford PA. Molecular effects of calcium binding mutations in Marfan syndrome depend on domain context. Hum Mol Genet. 2000 Aug 12;9(13):1987–94.

Rashid MH, Rumbaugh K, Free in PMC, Passador L, Davies DG, Hamood AN, Iglewski BH, Kornberg A. Polyphosphate kinase is essential for biofilm development, quorum sensing, and virulence of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A. 2000 Aug 15;97(17):9636–41.

Jormsjo S, Ye S, Moritz J, Walter DH, Dimmeler S, Zeiher AM, Henney A, Hamsten A, Eriksson P. Allele-specific regulation of matrix metalloproteinase12 gene activity is associated with coronary artery luminal dimensions in diabetic patients with manifest coronary artery disease. Circ Res. 2000 May 12;86(9):998–1003.

Talas U, Dunlop J, Khalaf S, Leigh IM, Kelsell DP. Human elastase 1: evidence for expression in the skin and the identification of a frequent frameshift polymorphism. J Invest Dermatol. 2000 Jan;114(1):165–70.

Horwitz M, Benson KF, Person RE, Aprikyan AG, Dale DC. Mutations in ELA2, encoding neutrophil elastase, define a 21-day biological clock in cyclic haematopoiesis. Nat Genet. 1999 Dec;23(4):433–6.

Griffin MD, Torres VE, Grande JP, Kumar R. Vascular expression of polycystin. J Am Soc Nephrol. 1997 Apr;8(4):616–26.

Existem referências para a relação entre mutações na elastase de neutrófilos e neutropenia – um baixo nível de células brancas do sangue chamadas neutrófilos. Para melhor compreender o assunto, podemos procurar por elastase no banco de dados de doenças genéticas humanas:

Online Mendelian Inheritance in Man (OMIM™)

OMIM é um banco de dados de genes humanos e de desordens genéticas. Ele foi originalmente compilado por V. A. McKusick, M. Smith e colegas e publicado em artigos científicos. O NCBI tornou este banco de dados acessível pela *web* e introduziu conexões para outros arquivos de informações correlatas, incluindo bancos de dados de seqüências e a literatura médica. OMIM encontra-se agora bem integrado com o sistema de recuperação de informações ENTREZ do NCBI. OMIM Morbid Map é um banco de dados ligado ao OMIM, mas que trata de doenças genéticas e suas localizações cromossômicas.

A resposta para ELASTASE em uma pesquisa no OMIM descreve resultados que relacionam as mutações no gene com a neutropenia cíclica.

O conjunto de resultados que reunimos sobre a elastase daria suporte a pesquisas sobre este sistema; por exemplo, poderíamos mapear mutantes da elastase na estrutura da molécula para ver se obteríamos indícios da causa das neutropenias cíclicas.

O Sequence Retrieval System (SRS)

O SRS, originalmente desenvolvido por T. Etzold, é um sistema integrado para recuperar informações de muitos bancos de dados de seqüências diferentes e para utilizar as seqüências recuperadas como entradas de ferramentas analíticas, tais como programas para comparação e alinhamento de seqüências.

O SRS pode pesquisar um total de 141 bancos de dados de seqüências de proteínas e nucleotídeos, vias metabólicas, estruturas 3D e funções, genomas, doenças e informações sobre o fenótipo (ver Quadro). Estes incluem muitos bancos de dados pequenos, como os bancos de dados Prosite e Blocks, de motivos estruturais de proteínas, bancos de dados de fatores de transcrição e bancos de dados especializados em certos agentes patogênicos.

Algumas categorias de bancos de dados que podem ser pesquisados a partir do SRS

Seqüência de nucleotídeos	Literatura
Uniprot	Mapeamento
Função das proteínas	Estrutura de proteínas
Enzimas	Vias metabólicas
Mutação, SNP	Ontologias de genes

Além do número e da variedade de bancos de dados aos quais o SRS oferece acesso, ele também disponibiliza conexões robustas entre os bancos de dados e fluidez na execução de aplicativos. Uma pesquisa em um único componente de um banco de dados pode ser estendida a uma pesquisa na rede inteira; ou seja, entradas em todos os bancos de dados relativas a uma dada proteína podem ser facilmente encontradas. Pesquisas por similaridade e alinhamentos podem ser executadas diretamente, sem a necessidade de armazenar as respostas em um arquivo intermediário.

Em uma sessão do SRS, você começa selecionando um ou mais bancos de dados para serem utilizados na pesquisa. Os bancos de dados são agrupados por categorias, como relacionados a seqüências nucleotídicas, relacionados a proteínas, etc. Então você pode entrar com um conjunto de termos de consulta. Assim como no ENTREZ, você deve procurar pelos termos em todos os campos, ou associar termos a categorias. O programa responderá com um conjunto de entradas que contenham os termos que você informou. Como opções para prosseguir com as consultas, pode-se:

1. Examinar uma das seqüências identificadas entrando no arquivo retornado como resultado.
2. Selecionar uma ou mais das seqüências identificadas e procurar por seqüências relacionadas em outros bancos de dados.
3. Executar uma aplicação, como uma predição da estrutura secundária ou um alinhamento múltiplo de seqüências.

Outras opções na página de resultados permitem que você crie e baixe em seu próprio computador relatórios sobre os alinhamentos selecionados. Isso pode ser simplesmente uma listagem das seqüências, ou o resultado de uma análise mais complexa dos resultados. Ao se aplicar o programa de alinhamento múltiplo de seqüências como o CLUSTAL-W aos resultados, obtém-se um alinhamento como o mostrado na Gravura III.

O Protein Information Resource (PIR)

O PIR é uma combinação efetiva de um banco de dados cuidadosamente organizado, um *software* de acesso à recuperação de informações e uma bancada para investigação de seqüências. O PIR também produz o Integrated Environment for Sequence Analysis (IESA) (ou Ambiente Integrado de Análise de Seqüências). Pense neste ambiente como um pacote de análise posicionado no topo de um sistema de recuperação de informações. Suas funcionalidades incluem navegar, pesquisar, analisar similaridades e conexões com outros bancos de dados. Os usuários podem:

- Navegar com base em anotações.
- Pesquisar em campos de texto selecionados por diferentes anotações, como, por exemplo, superfamília, família, título, espécie, grupo taxonômico, palavras-chave e domínios.
- Analisar seqüências utilizando o BLAST ou FASTA, pesquisar por padrão, alinhamento múltiplo.
- Fazer pesquisas por alinhamento global e por domínios, bem como pesquisas classificadas de acordo com anotação.
- Ver estatísticas para superfamília, família, título, espécies, grupo taxonômico, palavras-chave, domínios e características.
- Ver conexões para outros bancos de dados, incluindo PDB, COG, KEGG, WIT e BRENDA.
- Selecionar grupos de seqüências específicos, tais como o genoma humano, de camundongo, de levedura e *E. coli*.

As URLs para pesquisa no PIR por meio de expressões textuais são:
Nos Estados Unidos: `http://pir.georgetown.edu/pirwww/search/textsearch.shtml`
Na Europa: `http://mips.gsf.de/`

Uma característica do sistema internacional PIR é a pesquisa por um peptídeo específico. Olhando para o alinhamento de elastases de mamíferos na Gravura III, notamos, nas posições 220 a 228, um motivo conservado: a maioria das seqüências contém CNGDSGGPLN. No PIR, você pode selecionar PEPTIDE MATCH e pesquisar por combinações exatas para a subseqüência CNGDSGGPLN, obtendo uma lista com 84 resultados.

Retornando à tabela do alinhamento (Gravura III), em algumas moléculas observam-se variações no padrão. A pesquisa mais geral com o PATTERN SEARCH pelo padrão C[RNQF]GDSG[GS]PL[HNV], no qual [XYZ] significa uma posição contendo ou X, ou Y ou Z, obteria todas as elastases de mamíferos no alinhamento, mais um total de 320 seqüências ao todo. Mesmo essas não são todas as elastases homólogas no banco de dados, pois se pode fazer uma pesquisa PSI-BLAST por quaisquer das seqüências, ou, permanecendo estritamente dentro do PIR, consultar por elastases no banco de dados PIRSF. O padrão combina com 21 famílias, e todas são serinas proteinases.

Nós estamos bem no caminho para gerar uma lista completa de homólogos.

ExPASy – Expert Protein Analysis System

O ExPASy é um sistema de recuperação e análise de informação do Swiss Institute of Bioinformatics, o qual, em colaboração como o European Bioinformatics Institute, também produz os bancos de dados de seqüências de proteínas EMBL Data Library e TrEMBL. O TrEMBL contém traduções de seqüências de nucleotídeos do EMBL Data Library ainda não integradas completamente ao SWISS-PROT.

Abrindo a página principal do ExPASy (`http://www.expasy.org`) e selecionando SWISS-PROT e TrEMBL, obtém-se acesso a um conjunto de ferramentas de recuperação de informação, incluindo uma conexão para o SRS. Ainda há a opção de pesquisar o SWISS-PROT diretamente. Se selecionarmos FULL TEXT SEARCH e pesquisarmos o SWISS-PROT com o termo ELASTASE apenas, encontramos ELNE_HUMAN, o objetivo verdadeiro de nossa pesquisa, e aproximadamente 150 outros resultados, incluindo muitos inibidores. Uma homóloga da elastase encontrada é a do verme nematódeo causador da esquistossomose, o *Schistosoma mansoni*: CERC_SCHMA. Ambas as seqüências são precursoras; no alinhamento a seguir dessas duas seqüências, as letras maiúsculas indicam a enzima madura:

```
CERC_SCHMA  --msnrwrfvvvvtlftycltfervstwlIRSGEPVQHPAEFPFIAFLTTER-TMCTGSL 57
ELNE_HUMAN  mtlgrrlaclflacvlpalllggtalaseIVGGR-RARPHAWPFMVSLQLRGGHFCGATL 59
             :..*    :.:. ::.  *    . : * .*.    :*  :**:. *  .  :* .:*

CERC_SCHMA  VSTRAVLTAGHCVCSPLPVIRVSFLTLRNGDQQGIHHQPSGVKVAPGYMPSCMSARQRRP 117
ELNE_HUMAN  IAPNFVMSAAHCVAN—VNVRAVRVVLGAHNLSRREP—TRQVFAVQRIFENGYDP      111
            ::.. *::*.***..  :.*  : : * ::  :::*   .  : . : .  *

CERC_SCHMA  IAQTLSGFDIAIVMLAQMVNLQSGIRVISLPQPSDIPPPGTGVFIVGYGRDDNDRDPSRK 177
ELNE_HUMAN  VNLLN--DIVILQLNGSATINANVQVAQLPAQGRRLGNGVQCLAMGWGLLGRNRG—    164
            :      **.*: *   ..:::.::*  .**  .    *. : :*:*  ..:*.

CERC_SCHMA  NGGILKKGRATIMECRHATNGNPICVKAGQNFGQLPAPGDSGGPLLPS-LQGPVLGVVSH 236
ELNE_HUMAN  IASVLQELNVTVVTS-LCRRSNVCTLVRGRQAG—VCFGDSGSPLVCNGLIHGIASFVRG  221
            ..:*::  ..*:: .   . ..*  :  *:: * .  ****.**: . *    : ..*

CERC_SCHMA  GVTLPNLPDIIVEYASVARMLDFVRSNI----------------  264
ELNE_HUMAN  GCASGLYPDAFAPVAQFVNWIDSIIQRSEDNPCPHPRDPDPASRTH 267
            * :   ** :. *.... :* : ..
```

A estrutura da elastase de neutrófilos humana é conhecida da cristalografia por difração de raios X, mas a do esquistossomo não.

Um dos recursos exclusivos ao servidor do ExPASy é a conexão para o SWISS-MODEL, um servidor da *web* automático para a construção de modelos estruturais por homologia. Iniciando o SWISS-MODEL e escolhendo a opção FIRST APPROACH MODE (a mais simples), podemos simplesmente entrar com o código CERC_SCHMA do SWISS-PROT e iniciar a aplicação. A construção de um modelo por homologia não é uma operação trivial. Portanto, o trabalho é realizado em uma infra-estrutura computacional desconectada da *web* (*off-line*), e os resultados são enviados por correio eletrônico.

Discutiremos um pouco mais do SWISS-MODEL no Capítulo 5.

Ensembl

O Ensembl (`http://www.ensembl.org`) é planejado para ser a fonte universal de informação sobre o genoma humano. Os objetivos são coletar e anotar toda a informação disponível sobre seqüências de DNA humano, ligando-a à seqüência do genoma de referência e tornando-a acessível aos muitos cientistas que abordarão os dados sob os mais diferentes pontos de vista e diferentes necessidades. Para alcançar esse objetivo, além de coletar e organizar as informações, muitos esforços foram dirigidos para o desenvolvimento da infra-estrutura computacional. Convenções apropriadas de nomenclaturas estão estabelecidas: não é trivial planejar um esquema para manter identificadores estáveis frente a dados que estarão passando não somente por um crescimento, como também por revisão. O resultado mais visível desses esforços é o sítio da *web*, muito rico em recursos, tanto para a navegação quanto para a descrição de detalhes dos genomas.

O Ensembl é um projeto em conjunto do EBI e do Sanger Centre; os participantes incluem E. Birney, M. Clamp, T. Cox e T. J. P. Hubbard. Contudo, o Ensembl é organizado como um projeto aberto, o que estimula contribuições externas. Quase todos os leitores, até os mais ingênuos, devem reconhecer as grandes demandas que isso exigirá dos procedimentos de controle de qualidade.

Os dados coletados no Ensembl incluem genes, SNPs, repetições e homologias. Os genes podem ser conhecidos experimentalmente ou deduzidos a partir de uma seqüência. Como o apoio experimental para anotações de genes humanos é muito incerto, o Ensembl apresenta as evidências que corroboram a identificação de cada gene. Uma série de conexões para outros bancos de dados contendo informações relacionadas, como OMIM (Online Mendelian Inheritance in Man), ou bancos de dados de expressão de genes, amplia a informação acessível.

O Ensembl é estruturado em torno da seqüência do genoma humano. Os usuários podem identificar regiões por meio de muitos tipos de consultas ou pesquisas:

- Pesquisas BLAST sobre uma seqüência ou fragmento
- Navegação – começa no nível do cromossomo e então aplica o efeito zum
- Nome do gene
- Relações com doenças, via OMIM
- ENSEMBL ID, se o usuário o conhece
- Pesquisa de texto geral

Uma busca textual no Ensembl por BRCA1 produziu a página exibida na Figura 3.2, mostrando a região ao redor do *locus* BRCA1. O quadro da parte superior mostra uma região de 1×10^6 bases (um megabase) de DNA mapeada entre as bandas q21.2 e q21.31 do cromossomo 17. O quadro ainda descreve marcadores e genes anotados. O quadro inferior mostra uma visão mais detalhada.

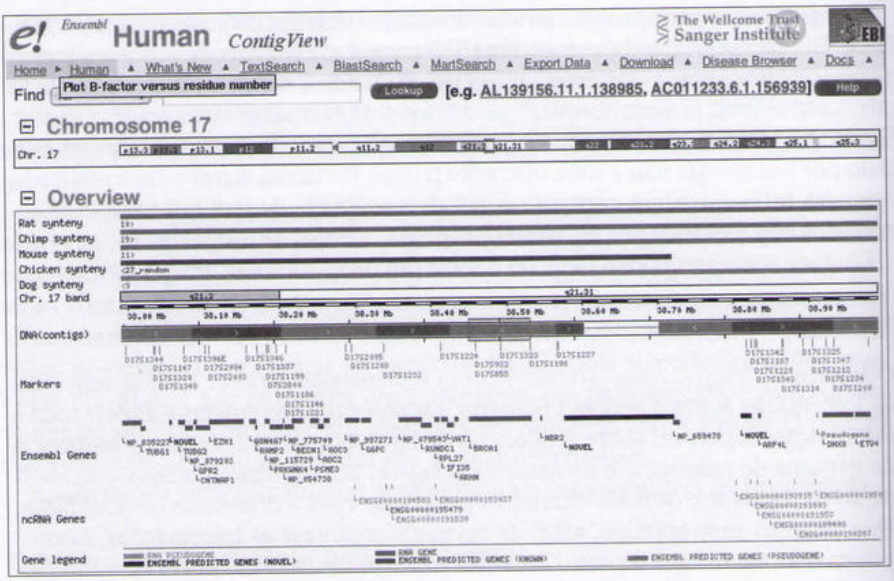


Figura 3.2 Uma parte da página do Ensembl mostrando uma região do genoma humano em torno do gene BRCA1.

Note que os painéis de controle entre os dois quadros permitem navegação e um zum variado. O quadro inferior mostra uma região de 0,1 megabase, descrevendo, portanto, muito mais detalhes, incluindo a estrutura detalhada do gene BRCA1 e dos SNPs observados.

Para onde vamos a partir daqui?

Visitamos somente alguns poucos bancos de dados de biologia molecular acessíveis na *web*. Em curto prazo, os leitores explorarão estes sítios e outros, e ficarão familiarizados não somente com os conteúdos da *web*, mas também com o seu dinamismo – a aparição e o desaparecimento de conexões e sítios. Existem várias metáforas biológicas para a *web* – como um ecossistema que está evoluindo, ou que está crescendo poluído, com sítios inativos e conexões para sítios inativos. Infelizmente, ainda não há um mecanismo eficiente para deterioração e reciclagem como no mundo orgânico!

Os bancos de dados estão desenvolvendo meios de intercomunicação mais efetivos, a ponto de as conexões, que antes eram limitadas, agora estarem formando uma rede de conexões mais coesa. Não está muito distante a existência de um único banco de dados de biologia molecular com inúmeras vias de acesso. Os cientistas poderão configurar o seu próprio acesso, selecionando partes da informação, criando "bancos de dados virtuais", adaptados às suas próprias necessidades.

Leituras recomendadas

A cada ano, a edição de janeiro da revista *Nucleic Acids Research* traz um conjunto de artigos sobre bancos de dados em biologia molecular. Esta é uma referência inestimável.

Bishop, M. J., *Genetics databases* (London: Academic, 1999). [Um compêndio de bancos de dados, acesso e análise.]

Zdobnov, E. M., Lopez, R., Apwelier, R. & Etzold, T. (2002), The EBI SRS server—new features, *Bioinformatics*, **18**, 1149-1150.

Exercícios, problemas e *weblemas*

Exercícios

3.1 Um banco de dados de veículos tem entradas para os seguintes tipos de veículos: bicicleta, triciclo, motocicleta, carro. Ele armazena somente as seguintes informações sobre cada entrada: (1) quantas rodas (um número) e (2) o meio de propulsão = humano ou motor. Para cada par de veículos possível, desenvolva uma combinação lógica de termos de consulta referentes ao valor exato ou ao intervalo do número de rodas, e ao meio de propulsão, que retornará os dois veículos selecionados e nenhum outro.

3.2 O Quadro da página 164 mostra a entrada do NCBI para a proteína ELASTASE 1 PRECURSOR humana. Em uma fotocópia desta página, indique quais itens são (a) puramente para o gerenciamento interno do banco, (b) dados periféricos, como referências bibliográficas, (c) os resultados de medidas experimentais, (d) informações inferidas a partir de medidas experimentais.

3.3 Escreva um programa em PERL para extrair a seqüência de aminoácidos de uma entrada no banco de dados ENTREZ de seqüências de proteínas, como mostrado no Quadro da página 164, e a converta no formato FASTA.

3.4 Compare o arquivo recuperado por uma pesquisa no NCBI por HUMAN ELASTASE no banco de dados de proteínas (página 164) e no banco de dados de nucleotídeos (página 165). Em fotocópias destas duas páginas, destaque com uma caneta marca-texto todos os itens que os dois arquivos têm em comum.

Weblemas

3.1 Obtenha a seqüência completa da entrada do SWISS-PROT para BOVINE PANCREATIC TRYPSIN INHIBITOR (e *não* PANCREATIC SECRETORY TRYPSIN INHIBITOR) e a entrada completa do PIR para esta proteína. Que informação cada um deles tem que o outro não tem?

3.2 Encontre uma lista de sítios espelho oficiais e não-oficiais do Protein Data Bank. Qual é o mais próximo de você?

3.3 Encontre todas as estruturas da mioglobina de baleias Sperm no Protein Data Bank e desenhe um histograma de suas datas de depósito.

3.4 Encontre estruturas de proteínas determinadas por Peter Hudson, sozinho ou com colegas.

3.5 Projete uma expressão (combinação de palavras-chave) de pesquisa para utilizar na ferramenta *Site Search* do Protein Data Bank que retorne estruturas da tioredoxina de *E. coli*, mas *não* estruturas de nucleases de estafilococos.

3.6 Qual fração das estruturas determinadas por cristalografia por difração de raios X depositadas no Protein Data Bank contém também arquivos com os fatores de estrutura depositados?

3.7 A entrada 8XIA do Protein Data Bank contém a estrutura de um monômero da D-Xilose isomerase de *Streptomyces rubiginosus*. Qual é a sua provável estrutura quaternária? Como era a geometria do conjunto correspondendo à provável estrutura quarternária derivada das coordenadas na entrada?

3.8 Encontre os vizinhos estruturais no Protein Data Bank para a entrada 2TRX (tioredoxina de *E. coli*.) de acordo com SCOP, CATH, FSSP e CE. Quais estruturas, se alguma, são consideradas vizinhas da 2TRX por todas essas classificações? Quais estruturas são consideradas vizinhas estruturais em algumas, mas não em todas as classificações?

3.9 Por que uma pesquisa por HUMAN ELASTASE no ENTREZ, na categoria de proteínas, retorna uma tRNA sintetase?

3.10 O Quadro da página 164 contém a seqüência de aminoácidos da ELASTASE 1 PRECURSOR humana. Quais as diferenças que existem entre a seqüência dessa proteína e a proteína madura?

3.11 Qual é a relação entre as seqüências de ELASTASE recuperadas de pesquisas no NCBI e no PIR?

3.12 Usando o SWISS-PROT diretamente, ou o SRS, recupere a entrada SWISS-PROT para HUMAN ELASTASE. Que informações este arquivo contém que não aparecem (a) na estrada correspondente do ENTREZ (proteínas) e (b) na entrada corresponde no PIR?

3.13 Que homólogas da NEUTROPHIL ELASTASE humana podem ser identificadas por PSI-BLAST?

3.14 Procure por estruturas de ELASTASES utilizando os recursos de pesquisa do Protein Data Bank. Compare os resultados com aqueles obtidos com o ENTREZ, descritos no texto.

3.15 Qual gene de *C. elegans* codifica uma proteína com seqüência similar à ELASTASE humana?

3.16 Qual é a localização cromossômica do gene humano para a GLUCOSE-6-PHOSPHATE DEHYDORGENASE (glicose-6-fosfato desidrogenase)?

3.17 Pseudogenes em eucariotos podem ser classificados entre aqueles que se originaram por duplicação de genes e divergência e aqueles reinseridos no genoma a partir de mRNA por um retrovírus, chamados pseudogenes *processados*. Pseudogenes processados podem ser identificados pela ausência de íntrons. Quais dos pseudogenes, se houver algum, no grupo de genes de globina humana são pseudogenes processados?

3.18 Análises genéticas preliminares durante o isolamento do gene associado com a fibrose cística localizaram-no entre o oncogene MET e a RFLP D7S8. Estimou-se então que esta região continha de 1 a 2 milhões pb e deveria conter de 100 a 200 genes. (a) Quantos pares de base esta região realmente possui? (b) Quantos genes expressos acredita-se que esta região agora contém?

3.19 O gene da síndrome Berardinelli-Seip foi inicialmente localizado entre dois marcadores na banda do cromossomo 11q13 – D11S4191 e D11S987. Quantos pares de base há no intervalo entre esses dois marcadores?

3.20 Existe algum banco de dados disponível na *web* que armazene especificamente informações estruturais e termodinâmicas sobre interações proteína-ácido nucléico?

3.21 O banco de dados do proteoma de levedura contém uma entrada para cdc6, a proteína que regula o início da replicação de DNA. (a) Em qual cromossomo

da levedura está o gene cdc6? (b) Por quais modificações pós-traducionais esta proteína passa até alcançar seu estado ativo maduro? (c) Quais são os parentes conhecidos mais próximos desta proteína em outras espécies? (d) Com quais outras proteínas se sabe que a cdc6 de levedura interage? (e) Qual é o efeito da distamicina A na atividade da cdc6 de levedura? (f) Qual é o efeito da actinomicina D na atividade da cdc6 de levedura?

CAPÍTULO 4

Alinhamentos e Árvores Filogenéticas

Sumário do capítulo

Introdução ao alinhamento de seqüências 178

O gráfico de pontos 180

Gráficos de pontos e alinhamentos de seqüências 185

Medidas de similaridade de seqüências 191

Esquemas de pontuação 191

Calculando o alinhamento de duas seqüências 195

Variações e generalizações 195

Métodos aproximados para a análise rápida de bancos de dados 196

O algoritmo de programação dinâmica para o alinhamento ótimo de seqüências par a par 196

A significância de alinhamentos 202

Alinhamento múltiplo de seqüências 206

Aplicações de alinhamentos múltiplos de seqüências em consultas a banco de dados 208

Perfis 209

PSI-BLAST 211

Modelos ocultos de Markov 213

Filogenia 218

Árvores filogenéticas 223

Métodos de agrupamento 225

Métodos cladísticos 226

O problema da variação das taxas de evolução 227

Considerações computacionais 228

Leituras recomendadas 229

Exercícios, problemas e weblemas 230

Objetivos do aprendizado

1. Entender o conceito de alinhamento de seqüências: a atribuição de correspondências entre pares de resíduos.
2. Saber como construir e interpretar gráficos de pontos e a relação entre gráficos de pontos e alinhamentos.
3. Ser capaz de definir e distinguir as distâncias de Hamming e de Levenshtein como medidas de similaridade de seqüências de caracteres.
4. Entender as bases dos esquemas de pontuação para alinhamentos de caracteres, incluindo matrizes de substituição e penalidades para lacunas.
5. Perceber a diferença entre os alinhamentos global e local e entender o uso de métodos aproximados para consultas rápidas a bancos de dados.
6. Entender o significado do escore Z e saber como interpretar valores P e valores E advindos de consultas a bancos de dados.
7. Ser capaz de interpretar alinhamentos múltiplos de seqüências de aminoácidos e fazer inferências sobre estruturas de proteínas a partir de alinhamentos múltiplos.
8. Ser capaz de definir e distinguir os conceitos de homologia, similaridade, agrupamento e filogenia.
9. Tornar-se perito no uso do PSI-BLAST e de programas correlacionados.
10. Perceber a utilidade de métodos de perfis e modelos ocultos de Markov na consulta a bancos de dados.
11. Entender o conteúdo e a significância de árvores filogenéticas e os métodos disponíveis para derivá-las, incluindo máxima parcimônia e máxima verossimilhança; saber a função de um grupo externo e como usá-lo na derivação de uma árvore filogenética.

Introdução ao alinhamento de seqüências

Dadas duas ou mais seqüências, nós inicialmente queremos:

- medir as suas similaridades
- determinar as correspondências entre pares de resíduos
- observar padrões de conservação e variabilidade
- inferir relações evolucionárias

Se pudermos fazer isso, estaremos em uma posição confortável para procurar por seqüências relacionadas em bancos de dados. Uma aplicação relevante é na anotação de genomas, envolvendo a atribuição de estrutura e função ao maior número possível de genes.

Como podemos definir uma medida quantitativa de similaridade entre seqüências? Para compararmos os nucleotídeos ou aminoácidos que aparecem em posições correspondentes em duas ou mais seqüências, precisamos primeiro determinar estas correspondências. O *alinhamento de seqüências é a identificação das correspondências entre pares de resíduos*. Ele é *a* ferramenta básica da bioinformática.

Qualquer atribuição de correspondência que preserve a ordem dos resíduos dentro da seqüência é um alinhamento. Lacunas (*gaps*) podem ser introduzidas.

Dadas duas seqüências de texto: primeira seqüência = `a b c d e`
segunda seqüência = `a c d e f`

um alinhamento razoável seria:

```
a b c d e -
a - c d e f
```

Nós precisamos então definir critérios para que um algoritmo possa escolher o *melhor* alinhamento. Para as seqüências `gctgaacg` e `ctataatc`, temos:

Um alinhamento sem sentido:

```
- - - - - - - g c t g a a c g
c t a t a a t c - - - - - - - -
```

Um alinhamento sem lacunas:

```
g c t g a a c g
c t a t a a t c
```

Um alinhamento com lacunas:

```
g c t g a - a - - c g
- - c t - a t a a t c
```

E mais um:

```
g c t g - a a - c g
- c t a t a a t c -
```

Muitos leitores considerariam o último destes alinhamentos o melhor dos quatro. Para decidir se esta é a melhor de *todas* as possibilidades, precisamos de um meio para examinar sistematicamente todos os alinhamentos possíveis. Então, precisamos computar escores refletindo a qualidade de cada alinhamento possível e identificar o alinhamento com o escore ótimo. O alinhamento ótimo pode não ser único: vários alinhamentos diferentes podem ter o mesmo valor de escore ótimo. Além disso, pequenas variações no esquema de escores podem modificar o ranqueamento dos alinhamentos, fazendo com que um alinhamento inesperado apareça como o melhor.

Estes exemplos ilustram os **alinhamentos de seqüências par a par**. Entretanto, freqüentemente, podemos encontrar muitas famílias de seqüências similares pela identificação de homólogas em espécies diferentes. Um alinhamento simultâneo de mais de duas seqüências é chamado de **alinhamento múltiplo de seqüências**. Os alinhamentos múltiplos de seqüências são muito mais informativos do que os alinhamentos de seqüências par a par em termos de revelar padrões de conservação.

O gráfico de pontos

O gráfico de pontos é uma forma simples de visualização que fornece um panorama da similaridade entre duas seqüências (par a par). Porém, a sua relação com os alinhamentos é menos óbvia.

O gráfico de pontos é uma tabela ou matriz. As linhas correspondem aos resíduos de uma seqüência, e as colunas, aos resíduos da outra seqüência. Na sua forma mais simples, as posições no gráfico de pontos são deixadas em branco se os resíduos forem diferentes, e preenchidas se eles forem iguais. Segmentos de resíduos similares são mostrados como linhas diagonais no sentido do canto superior esquerdo para o canto inferior direito (direção noroeste-sudeste).

Exemplo 4.1

Gráfico de pontos mostrando a identidade entre o nome abreviado (DOROTHYHODGKIN) e o nome completo (DOROTHYCROWFOOTHODGKIN) de uma famosa cristalógrafa de proteínas:

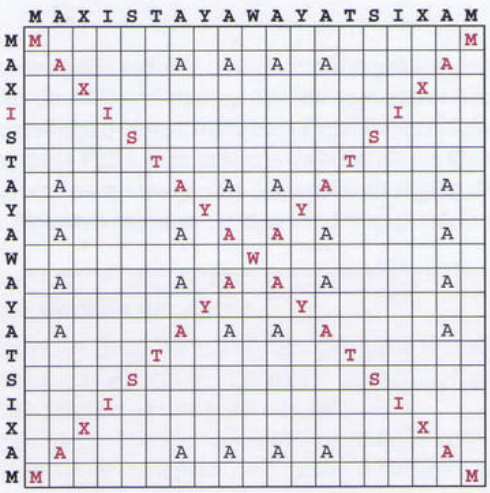
As letras correspondentes aos pareamentos idênticos *isolados* são mostradas como caracteres simples. As regiões com pareamentos idênticos longos, mostradas em vermelho, são o primeiro e o último nomes, DOROTHY e HODGKIN. Regiões de pareamento idêntico menor, como a OTH de dorOTHy e crowfoOTHodgkin, ou a RO de doROthy e cROwfoot, são consideradas ruído.

Exemplo 4.2

Gráfico de pontos mostrando identidades entre uma seqüência repetitiva (ABRACADABRACADABRA) e ela mesma. As repetições aparecem em várias diagonais secundárias, paralelas à diagonal principal.

Exemplo 4.3

Gráfico de pontos mostrando identidades entre a seqüência palindrômica MAX I STAY AWAY AT SIX AM e ela mesma. O palíndromo se revela como uma região de identidade *perpendicular* à diagonal principal.



Este não é apenas um jogo de palavras – regiões do DNA reconhecidas por reguladores de transcrição ou por enzimas de restrição possuem seqüências relacionadas com palíndromos, entrecruzadas nas fitas da cadeia de DNA:

→

Exemplo 4.3 (*continuação*)

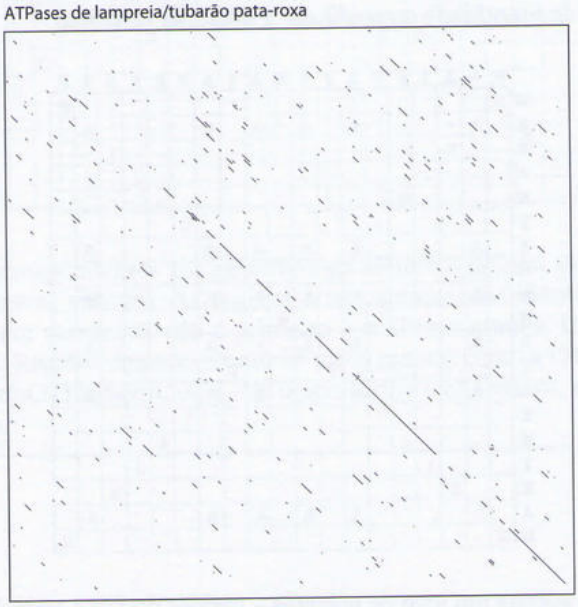
Sítio de reconhecimento da EcoRI: `GAATTC`
`CTTAAG`

Em cada fita, uma região é seguida pela sua seqüência complementar reversa (ver Exercício 4.9 e Problema 4.8). Regiões mais longas de DNA ou RNA contendo repetições invertidas como esta podem formar estruturas haste-alça. Além disso, alguns elementos de transposição em plantas contêm seqüências palindrômicas verdadeiras (aproximadas) – repetições invertidas de seqüências não-complementares, na mesma fita. O exemplo a seguir é encontrado no genoma do vírus anão do trigo: `ttttcgtgagtgcg-cggaggctttt`.

O gráfico de pontos permite uma representação pictórica rápida da relação entre duas seqüências. Características óbvias de similaridade se destacam. Por exemplo, o gráfico de pontos relacionando os genes da ATPase-6 mitocondrial de uma lampreia (*Petromyzon marinus*) e do tubarão pata-roxa (*Scyliorhinus canicula*) mostra que a similaridade entre as seqüências é fraca na porção inicial. Este gene codifica uma subunidade do complexo ATPase. Em humanos, mutações nesse gene causam a síndrome de Leigh, um distúrbio neurológico em crianças produzido pelos efeitos do metabolismo oxidante desregulado no cérebro durante o seu desenvolvimento.

Uma desvantagem do gráfico de pontos é que seu "alcance" na comparação entre seqüências distantemente relacionadas é baixo. Na análise de seqüências, deve-se sempre analisar o gráfico de pontos para se ter certeza de que algo óbvio não está sendo perdido, mas deve-se também estar preparado para a aplicação de ferramentas mais refinadas.

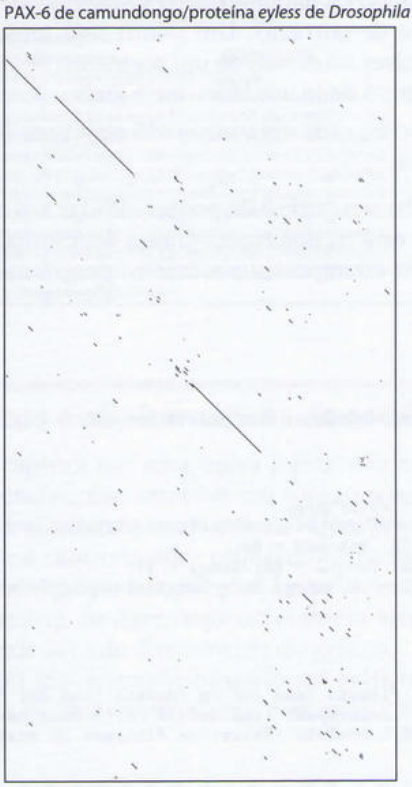
ATPases de lampreia/tubarão pata-roxa



Freqüentemente, regiões de similaridade são deslocadas, para aparecerem em diagonais paralelas, mas não colineares. Isso significa que inserções ou deleções

ocorreram nos segmentos entre as regiões similares. O gráfico de pontos relacionando a proteína PAX-6 de camundongo e a proteína *eyeless* de *Drosophila melanogaster* mostra três regiões estendidas de similaridade com diferentes comprimentos de seqüência entre elas. Duas dessas regiões estão próximas do início das seqüências e uma está no meio. A seção entre a segunda e a terceira região de similaridade é mais longa na seqüência de camundongo do que na seqüência de *Drosophila*.

PAX-6 de camundongo/proteína *eyless* de *Drosophila*



Filtrar os resultados pode reduzir o ruído em um gráfico de pontos. Na comparação das seqüências de ATPases, pontos não são mostrados a menos que eles estejam no centro de uma região consecutiva de 15 resíduos contendo pelo menos seis identidades. O programa PERL para gráficos de pontos (ver Quadro) permite que o usuário estipule valores para uma janela (extensão da região de resíduos consecutivos) e um limiar (número mínimo de identidades necessário dentro da janela).

Recursos na *web*: Gráficos de pontos

O programa Dotter de E. L. Sonnhammer calcula e representa gráficos de pontos. Ele permite que o usuário controle o cálculo e altere a aparência do gráfico pelo ajuste interativo dos parâmetros.

`http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html`

Para utilizar todas as ferramentas do Dotter, é preciso instalá-lo em seu computador.

Um sítio na web que oferece a geração interativa de gráficos de pontos é:

`http://myhits.isb-sib.ch/cgi-bin/dotlet`

Exemplo PERL 4.1 Um programa para desenhar gráficos de pontos

O programa mostrado logo abaixo lê:

1. Um título geral para o gráfico, que será impresso no topo do gráfico de saída (primeira linha do arquivo de entrada).
2. Dados especificando os parâmetros de filtragem *janela* e *limiar* (segunda linha do arquivo de entrada). Um ponto será mostrado no gráfico de pontos se ele estiver no centro de um segmento de resíduos de extensão *janela* e se o número de identidades for ≥ *limiar*.
3. As duas seqüências, cada uma iniciando com uma linha de título e terminando com um *.

O programa desenha um gráfico de pontos similar aos mostrados no texto. O arquivo de saída está na linguagem gráfica denominada PostScript™, que pode ser visualizada ou impressa por muitos programas.

```
#!/usr/bin/perl
#dotplot.pl -- reads two sequences and prints dotplot

# read input

$/ = "";
$_ = <DATA>; $_ =~ s/#(.*)\n/\n/g;
$_ =~ /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(.*)\n([A-Za-z\n]*)\*\s*\n(.*)\n([A-Za-z\n]*)\*/;
$title = $1; $nwind = $2; $thresh = $3;
$seqt1 = $4; $seq1 = $5; $seqt2 = $6; $seq2 = $7;
$seq1 =~ s/\n//g; $seq2 =~ s/\n//g; $n = length($seq1); $m = length($seq2);

# postscript header

print <<EOF;
%!PS-Adobe-
/s /stroke load def /l /lineto load def /m /moveto load def /r /rlineto load def
/n /newpath load def /c /closepath load def /f /fill load def
1.75 setlinewidth 30 30 translate /Helvetica findfont 20 scalefont setfont
EOF

#print matrix

$dx = 500.0/$n;  $mdx = -$dx; $dy = 500.0/$m;
if ($dy < $dx) {$dx = $dy;} $dy = $dx; $xmx = $n*$dx; $ymx = $m*$dx;
print "0 510 m ($title NWIND = $nwind) show\n";
printf "0 0 m 0 %9.2f l %9.2f %9.2f l %9.2f 0 l c s\n", $ymx,$xmx,$ymx,$xmx;

for ($k = $nwind - $m + 1; $k < $n - $nwind; $k++) {
    $i = $k; $j = 1; if ($k < 1) {$i = 1; $j = 2 - $k;}
    while ($i <= $n - $nwind && $j <= $m - $nwind) {
        $_ = (substr($seq1,$i -1,$nwind) ^ substr($seq2,$j -1,$nwind));
        $mismatch = ($_ =~ s/[^\x0]//g);
        if ($mismatch < $thresh) {
            $xl = ($i - 1)*$dx; $yb = ($m - $j)*$dy;
            printf "n %9.2f %9.2f m %9.2f 0 r 0 %9.2f r %9.2f 0 r c f\n",
                $xl,$yb,$dx,$dy,$mdx;
        }
        $i++; $j++;
    }
}
print "showpage\n";

__END__
ATPases lamprey / dogfish                    #TITLE
15 6                                         #WINDOW, THRESHOLD
Petromyzon marinus mitochondrion             #SEQUENCE 1
atgacactagatatctttgaccaatttacctccccaaca
atatttgggcttccactagcctgattagctatactagcccctagctta
atattagtttcacaaacaccaaaatttatcaaatctcgttatcacacacta
```

```
atatttgggcttccactagcctgattagctatactagcccctagctta
atattagtttcacaaacaccaaaatttatcaaatctcgttatcacacacta
cttacacccatcttaacatctattgccaaacaactctttcttccaataaac
caacaagggcataaatgagccttaatttgtatagcctctataatatttatc
ttaataattaatctttaggattattaccatatacttatacaccaactacc
caattatcaataaacataggattagcagtgccactatgactagctactgtc
ctcattgggttacaaaaaaaaccaacagaagccctagcccacttattacca
gaaggtaccccagcagcactcattcccatattaattatcattgaaactatt
agtctttttatccgacctatcgcccctaggagtccgactaaccgctaattta
acagctggtcacttacttatacaactagtttctataacaacctttgtaata
attcctgtcatttcaatttcaattattacctcactacttcttctatta
ctaacaattctggagttagctgttgctgtaatccaggcatatgtatttatt
ctacttttaactctttatctgcaagaaaacgttt*
Scyliorhinus canicula mitochondrion          #SEQUENCE 2
atgattataagctttttgatcaattcctaagtccctcctttctagga
atcccactaattgccctagctatttcaattccatgattaatatttccaacaccaacc
aatcgttgacttaataatcgattattaactcttcaagcatgatttattaaccgatttatt
tatcaactaatacaacccataaatttaggaggacataaatgagctatcttatttacagcc
ctaatattatttttaattaccatcaatcttctaggtctccttccatatactttttacgcct
acaactcaactttctcttaatatagcctttgccctgcccttatggcttacaactgtatta
attggtatatttaatcaaccaaccattgccctagggcacttattacctgaaggtacccca
acccctttagtaccagtactaatcattatcgaaaccatcagtttatttattcgaccatta
gccttaggagtccgattaacagccaacttaacagctggacatctccttatacaattaatc
gcaactgcggcctttgtccttttaactatataataccaaccgtggccttactaacctcccta
gtcctgttcctattgactattttagaagtggctgtagctataattcaagcatacgtattt
gtccttctttttaagcttatatctacaagaaaacgtataa*
```

Gráficos de pontos e alinhamentos de seqüências

O gráfico de pontos captura em uma única figura não apenas a similaridade geral entre duas seqüências, mas também um conjunto completo e a qualidade relativa dos diferentes alinhamentos possíveis. Qualquer caminho através do gráfico de pontos entre o canto superior esquerdo e o canto inferior direito, movendo-se em cada ponto apenas nas direções Leste, Sul e Sudeste, corresponde a um alinhamento possível. Se duas seqüências forem estreitamente relacionadas, o alinhamento pode ser lido diretamente do gráfico.

A Figura 4.1 mostra um exemplo baseado no gráfico de pontos Dorothy Hodgkin. Se a direção do "movimento" entre as celas sucessivas for diagonal,

Figura 4.1 Qualquer caminho através do gráfico de pontos, do canto superior esquerdo até o canto inferior direito, passa por uma sucessão de celas, *cada uma correspondendo a um par de posições, uma da linha e uma da coluna, que são correspondentes no alinhamento, ou que indica uma lacuna em uma das seqüências.* O caminho não precisa passar apenas por celas preenchidas por pontos. No entanto, quanto maior o número de celas preenchidas por pontos no caminho, maior o pareamento de resíduos idênticos no alinhamento.

dois pares de resíduos sucessivos aparecem no alinhamento sem uma inserção entre eles. Se a direção do movimento for horizontal, uma lacuna é introduzida na seqüência que indexa as linhas. Se a direção do movimento for vertical (nenhum neste exemplo), uma lacuna é introduzida na seqüência que indexa as colunas. Note que nenhum movimento pode ser realizado para cima ou para a esquerda, pois isso significaria alinhar diversos resíduos de uma seqüência com apenas um resíduo da outra. O caminho indicado pelas setas corresponde ao alinhamento evidente:

```
DOROTHY-------HODGKIN
DOROTHYCROWFOOTHODGKIN
```

Outra forma de imaginar um caminho através do gráfico de pontos é um *script de edição*, ou seja, a prescrição de uma série de operações que transformam a seqüência que indexa as colunas – a seqüência "horizontal" – na seqüência que indexa as linhas – a seqüência "vertical". Cada movimento nos diz qual operação deve ser executada – uma substituição, um inserção ou uma deleção. Quando o final do caminho é atingido, o efeito será a transformação de uma seqüência na outra. Em geral, várias seqüências de operações de edição diferentes podem converter uma seqüência de caracteres em outra com o mesmo número de passos, mas elas podem originar diferentes alinhamentos.

Deve ser enfatizado que, embora a seqüência de operações de edição derivada de um alinhamento ótimo *possa* corresponder a uma via evolucionária verdadeira, é impossível *provar* que ela de fato corresponda a esta via. Quanto maior a distância de edição, maior o número de vias evolucionárias plausíveis entre duas seqüências.

Exemplo 4.4 Gráficos de pontos e alinhamentos

Vamos comparar a aparência dos gráficos de pontos de pares de proteínas cujas relações são cada vez mais distantes. A Figura 4.2 mostra os gráficos de pontos das comparações da cisteína proteinase papaína do mamão (papaia) com quatro proteínas homólogas – o parente próximo, a actinidina da fruta quiuí e parentes mais distantes, a procatepsina L humana, a catepsina B humana e a estafilopaína (*staphopain*) de *Staphylococcus aureus*. Os alinhamentos das seqüências também são mostrados. Conforme as seqüências divergem progressivamente, torna-se mais e mais difícil determinar o alinhamento correto no gráfico de pontos. Os alinhamentos mostrados foram derivados das comparações das estruturas.

→

```
ALIGNMENT OF 9pap      and 2act
SCORE = 5324  NPOS = 219 NIDENT = 102  %IDENT = 46.58

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDR--
 | |||||  |||   | || || ||||||  | ||| ||  | |   ||||| || |
LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ

RSYGCNGGYPWSALQ-LVAQYGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYN
   || |||     |       ||     |||       |           |    |   |
NTRGCDGGYITDGFQFIINDGGINTEENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNN

QGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGP----NYILI
  ||      ||||| | |||  |  |  ||| ||||  ||||   ||||      |
EWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWIV

KNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
|||| | ||| || || |  |   | ||  |   ||||
KNSWDTTWGEEGYMRILRNVGGA-GTCGIATMPSYPVKY
```

PAPA_CARPA/ACTN_ACTCH

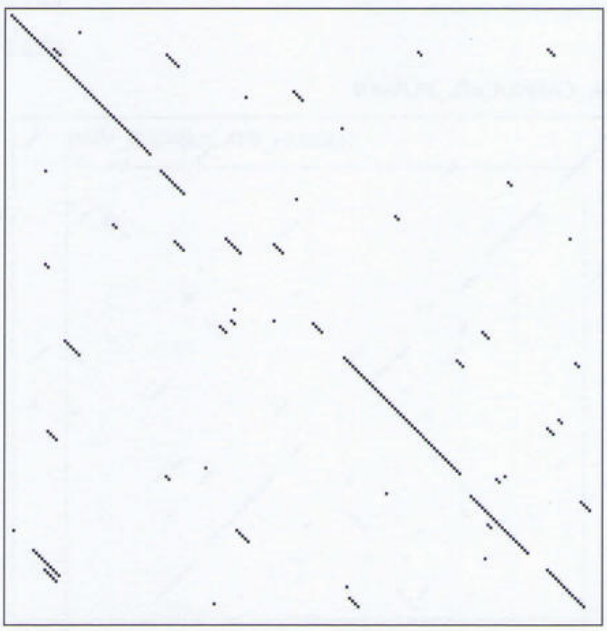


Figura 4.2a Alinhamento da papaína de mamão (papaia) e da actinidina da fruta quiuí, com o gráfico de pontos correspondente.

Exemplo 4.4 (*continuação*)

```
ALIGNMENT OF 9pap      and 1cjl

SCORE = 3214 NPOS = 220 NIDENT  = 81  %IDENT = 36.82

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCD--R
     ||| || |||||||| ||| |||||     ||      || |   ||| | ||
V----DWREKGYVTPVKNQGQCGSSWAFSATGALEGQMFRKTGRLISLSEQNLVDCSGPE

RSYGCNGGYPWSALQLVAQY-GIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYN
   |||||    | | |     |      ||||     |    |    |    |
GNEGCNGGLMDYAFQYVQDNGGLDSEESYPYEATEESCKYNPKYS-VANDAGFVDIPKQE

QGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGP--CGNKVDHAVAAVGYG---PNYIL
          | ||   |    | |  ||   |        || |  ||||     | |
KALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPDCSSEDMDHGVLVVGYGFESNKYWL

IKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
 |||||  ||  ||         |    ||      ||
VKNSWGEEWGMGGYVKMAKDRRN-H--CGIASAASYPTV-
```

PAPA_CARPA/CATL_HUMAN

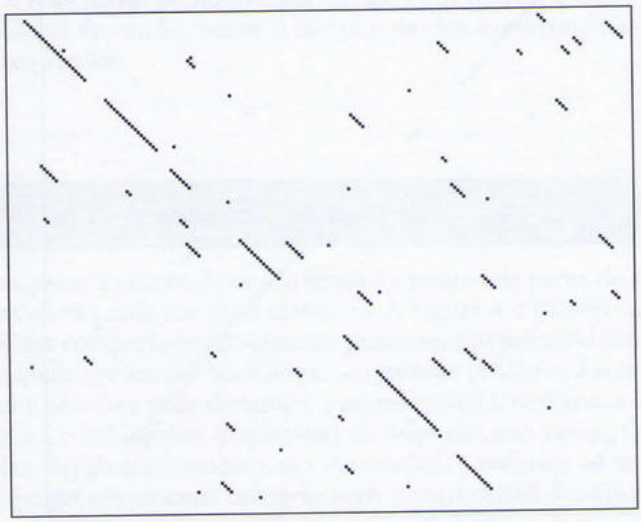


Figura 4.2b Alinhamento da papaína de mamão (papaia) e da procatepsina L humana, com o gráfico de pontos correspondente. Este gráfico de pontos mostra que existem diversas regiões de similaridade, mas seria difícil gerar um alinhamento de seqüências completo a partir dele.

```
ALIGNMENT OF 9pap     and 1huc

SCORE   =  2073  NPOS = 251  NIDENT  = 66  %IDENT = 26.29

IPEYVD-WRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLD-C-D
      | |          |||||||||| ||  |   | | |       |   ||  |
--DAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHTNVSVEVSAEDLLTCCGS

RRSYGCNGGYP------WSALQLVAQYGI--HYRN-TY-----P--YEGVQRYCRSREKG
    |||||||      |    ||       |     |          |    |
MCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGEGDT

PYAAK------TDGVRQVQPYNQGALLYSIANQPVSV-V-----LQ---AAGKDFQLYRG
|   |          |   |       |                          || ||
PKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKS

GIFVGPCGNKV-DHAVAAV--GY--GPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVCG
|      |     ||       |  | | |  ||| | || ||   | ||      |
GVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQ-DHCGIES

LYTSSFYPVKN
       |
EVVAGI-PRTD
```

PAPA_CARPA/CATB_HUMAN

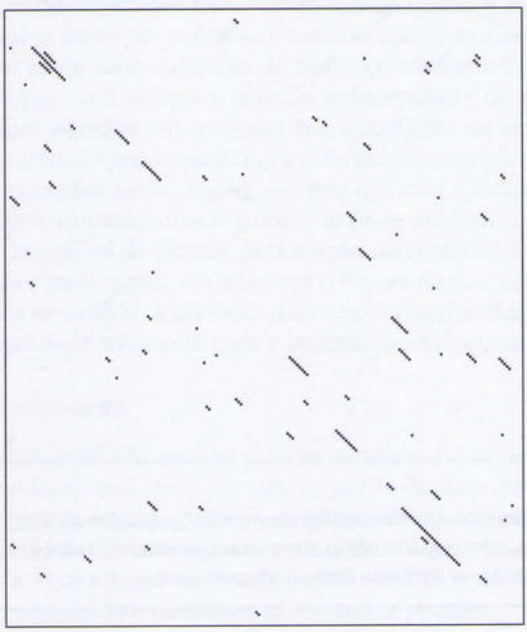


Figura 4.2c Alinhamento da papaína de mamão (papaia) e da catepsina B de fígado humana, com o gráfico de pontos correspondente. Note que, *tanto* no alinhamento das seqüências *quanto* no gráfico de pontos, a região de maior similaridade está localizada no início e no fim das seqüências, em relação à região central.

Exemplo 4.4 (*continuação*)

```
ALIGNMENT OF 9pap      and 1cv8

SCORE = -290  NPOS = 219  NIDENT  = 25 %IDENT = 11.42

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRS
                                                |     |
-----------------------------------------------EQYVNKLENFKIRE

YGCNGGYPWSALQLVAQYGIHYRNTYPYEGVQRYCRSREKG-PYAAKTDGVRQVQPY---
   | |                  | |  | | |       |          |
TQGNNGWCAGYTMSALLNATYNTNKYHAEAVMRFLHPNLQGQQFQFTGLTPREMIYFGQT

--NQGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPNYILIK
      ||           |     |       |       |    || | ||
QGRSPQLLNRMTTYNEVDNLTKNNKGIAIL-GSRVESRNGMHAGHAMAVVGNAKLNNGQE

NSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN-
         ||              |
VIIIWNPWDNGFMTQDAKNNVIPVSNGDHYQWYSSIYGY
```

PAPA_CARPA/STPA_STAAU

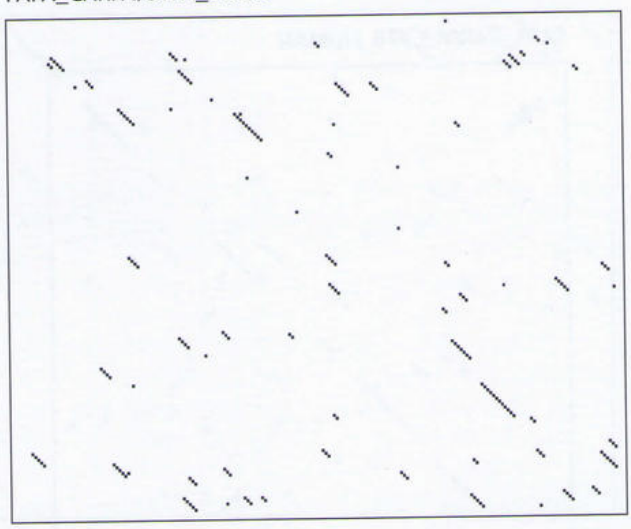


Figura 4.2d Alinhamento da papaína de mamão (papaia) e da estafilopaína (*staphopain*) de *S. aureus*, com o gráfico de pontos correspondente. O alinhamento destas duas seqüências não pode ser derivado deste gráfico de pontos.

Medidas de similaridade de seqüências

Para ir além do "alinhamento a olho" através de gráficos de pontos, é necessário definir medidas quantitativas de similaridade e de diferença entre seqüências.

Dadas duas seqüências, duas medidas da distância entre elas são:

(1) A **distância de Hamming**, definida entre duas seqüências de comprimentos iguais, é o número de posições com caracteres malpareados.

(2) A **distância de Levenshtein**, ou **distância de edição**, entre duas seqüências de comprimentos não necessariamente iguais, é o número mínimo de "operações de edição" necessário para converter uma seqüência na outra, onde uma operação de edição é uma deleção, inserção ou alteração de um único caractere em uma das seqüências. Uma dada seqüência de operações de edição leva a um alinhamento único, mas o contrário não é verdadeiro.

Por exemplo:

```
agtc        Distância de Hamming = 2
cgta

ag-tcc      Distância de Levenshtein = 3
cgctca
```

Para aplicações na biologia molecular, é preciso assumir que certas alterações têm maior probabilidade de ocorrer do que outras. Por exemplo, substituições de aminoácidos tendem a ser conservativas: a troca de um aminoácido por outro de tamanho ou propriedades físico-químicas similares é mais provável de ter ocorrido do que a troca por outro aminoácido com propriedades dissimilares. Ainda, a deleção de uma sucessão de bases contíguas ou de aminoácidos é um evento mais provável do que a deleção independente do mesmo número de bases ou de aminoácidos em posições não-contíguas na seqüência. Dessa forma, desejamos atribuir pesos variáveis a diferentes operações de edição. Um programa de computador pode, então, determinar não apenas a distância de edição, mas também alinhamentos ótimos. Ele pode atribuir um escore a cada caminho através do gráfico de pontos pela adição dos escores de cada passo individual. Para cada substituição, ele adiciona o escore da mutação, dependendo do par de resíduos envolvido. Para cada movimento horizontal ou vertical, ele adiciona uma penalidade adequada para a introdução da lacuna.

Esquemas de pontuação

Um sistema de pontuação (de escore) precisa considerar a substituição de resíduos, inserções ou deleções (uma inserção, do ponto de vista de uma seqüência, é uma deleção do ponto de vista da outra seqüência!). Deleções, ou lacunas em uma seqüência, terão um escore que dependerá da sua extensão.

As distâncias de Hamming e de Levenshtein medem a *dissimilaridade* entre duas seqüências: seqüências similares apresentam pequenas distâncias e seqüências dissimilares apresentam grandes distâncias. É comum na biologia molecular definir escores como medidas de *similaridade* entre seqüências. Dessa maneira, seqüências similares apresentam altos escores e seqüências dissimilares apresentam baixos escores. Estas formulações são equivalentes. Algoritmos para um alinhamento ótimo podem procurar tanto minimizar uma medida de dissimilaridade quanto maximizar uma função de pontuação.

Para seqüências de ácidos nucléicos, é comum utilizar um esquema simples para substituições: +1 para identidade e –1 para bases não-idênticas, ou um

Exemplo 4.5

Mutações do tipo *transição* (*purina* ↔ *purina* e *pirimidina* ↔ *pirimidina*, ou seja, a ↔ g e t ↔ c) são mais comuns do que *transversões* (*purina* ↔ *pirimidina*, ou seja, (a ou g) ↔ (t ou c)). Sugira uma matriz de substituição que reflita este fato.

Uma possibilidade é:

	a	g	t	c
a	20	10	5	5
g	10	20	5	5
t	5	5	20	10
c	5	5	10	20

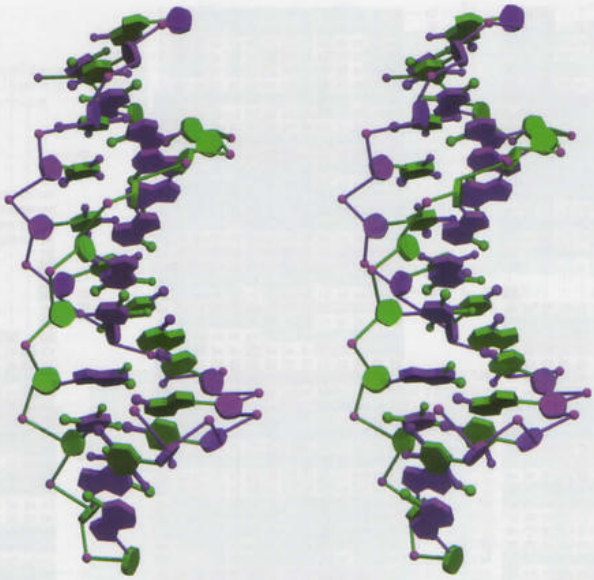
esquema mais complicado baseado na maior freqüência de mutações do tipo transições do que mutações transversas.

Para proteínas, uma variedade de esquemas de pontuação já foi proposta. Podemos agrupar os aminoácidos em classes de acordo com suas similaridades físico-químicas e atribuir +1 para pareamento (ou identidade) dentro de uma classe de resíduos e –1 para resíduos em classes diferentes. Podemos tentar desenvolver um escore de substituição mais preciso com base na combinação de propriedades dos aminoácidos. Ou, de modo alternativo, podemos tentar deixar as proteínas nos ensinarem um esquema de pontuação mais apropriado. M. O. Dayhoff foi a pioneira nesta área, colecionando estatísticas sobre a freqüência de substituições nas seqüências de proteínas até então conhecidas (década de 1960). Seus resultados foram utilizados durante muitos anos para pontuar alinhamentos. Eles foram substituídos por novas matrizes baseadas em conjuntos muito maiores de seqüências que se tornaram disponíveis.

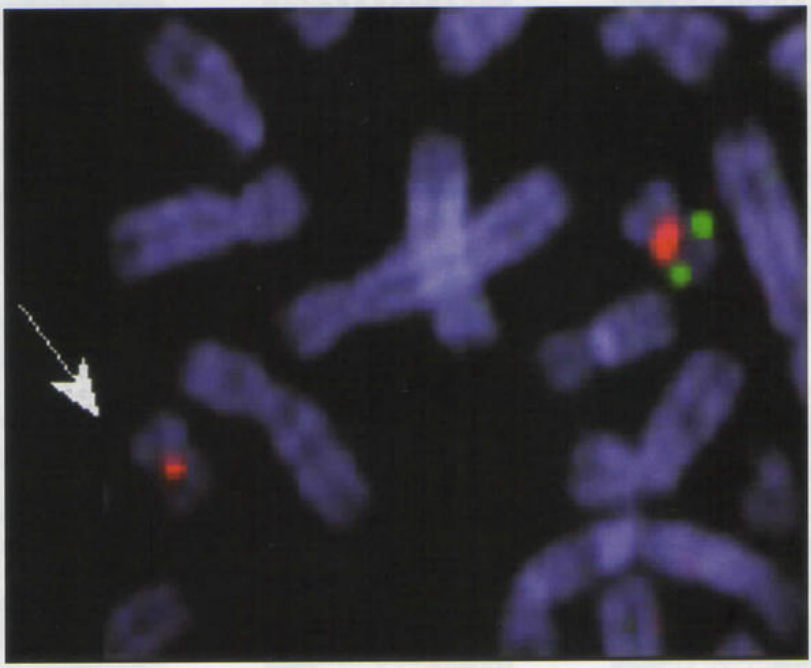
Derivação de matrizes de substituição

À medida que as seqüências divergem, mutações se acumulam. Para medir a probabilidade relativa de qualquer substituição em particular, por exemplo, serina → treonina, podemos contar o número de mudanças serina → treonina em pares de seqüências homólogas alinhadas. Poderíamos utilizar as freqüências relativas destas alterações para criar uma matriz de escores para as substituições. Uma mudança mais provável apresentará um escore maior do que uma mudança rara. Mas, e se houver múltiplas substituições em certas posições? Isso influenciará a estatística. Podemos evitar este problema restringindo nossas amostras a seqüências suficientemente similares de forma que possamos assumir que nenhuma posição foi alterada mais do que uma vez.

Uma medida da divergência de seqüências é o **PAM**: 1 PAM = 1 por cento de mutação aceita (1 Percent Accepted Mutation). Assim, duas seqüências separadas por 1 PAM têm 99% dos seus resíduos idênticos. Para pares de seqüências com nível de divergência de 1 PAM, é provável que não tenha ocorrido mais do que uma alteração em qualquer uma das suas posições. Com a análise de estatísticas de pares de seqüências próximas, como o exemplo acima, e a correção dos valores de acordo com a abundância relativa de diferentes aminoácidos, são produzidas as **matrizes de substituições 1 PAM**. Para produzir uma matriz apropriada para seqüências com maior divergência, podemos tomar potências dessa matriz. O nível PAM250, correspondente a aproximadamente 20% de identidade

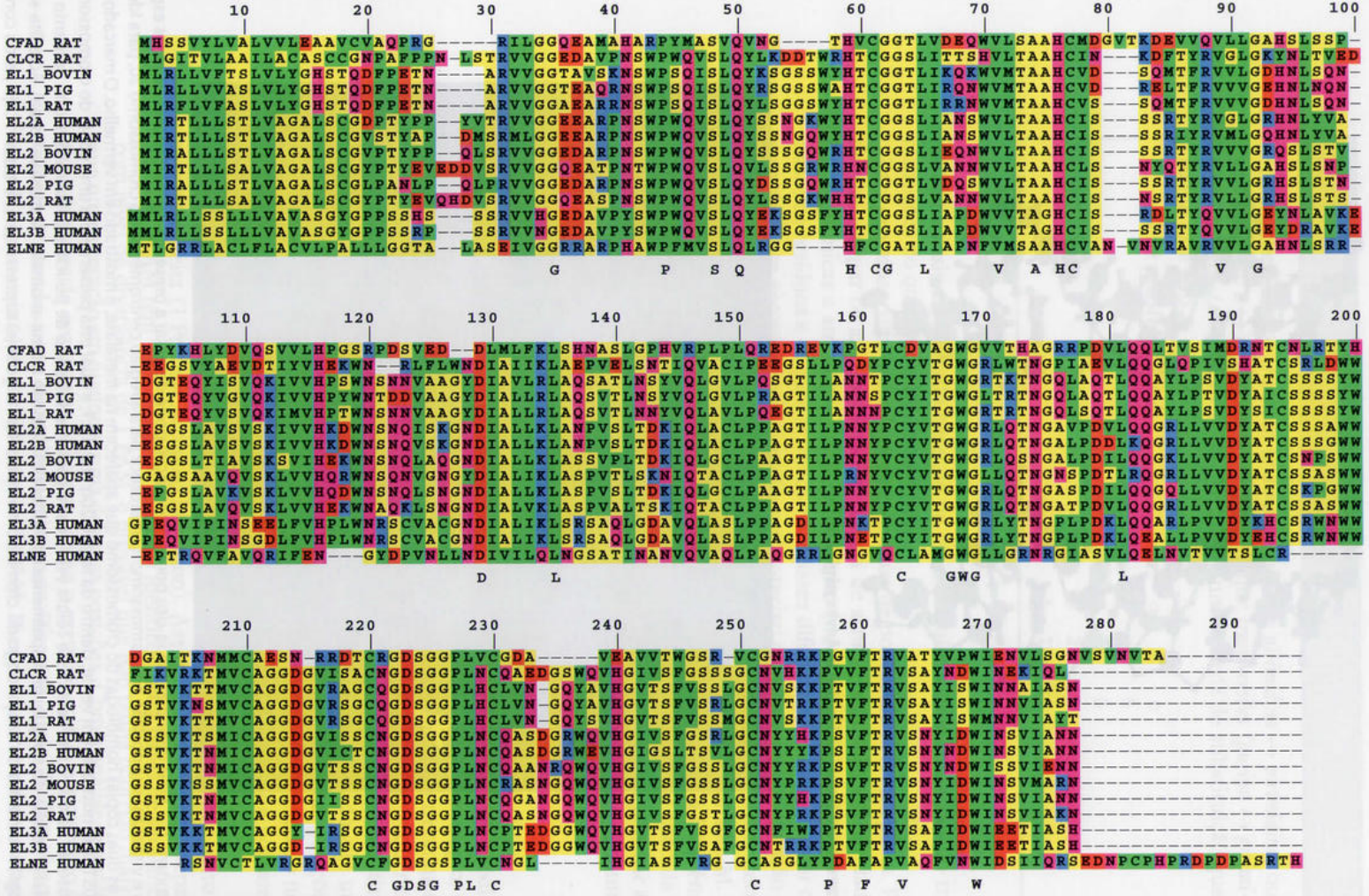


Gravura I Dupla hélice do DNA (ver página 26).

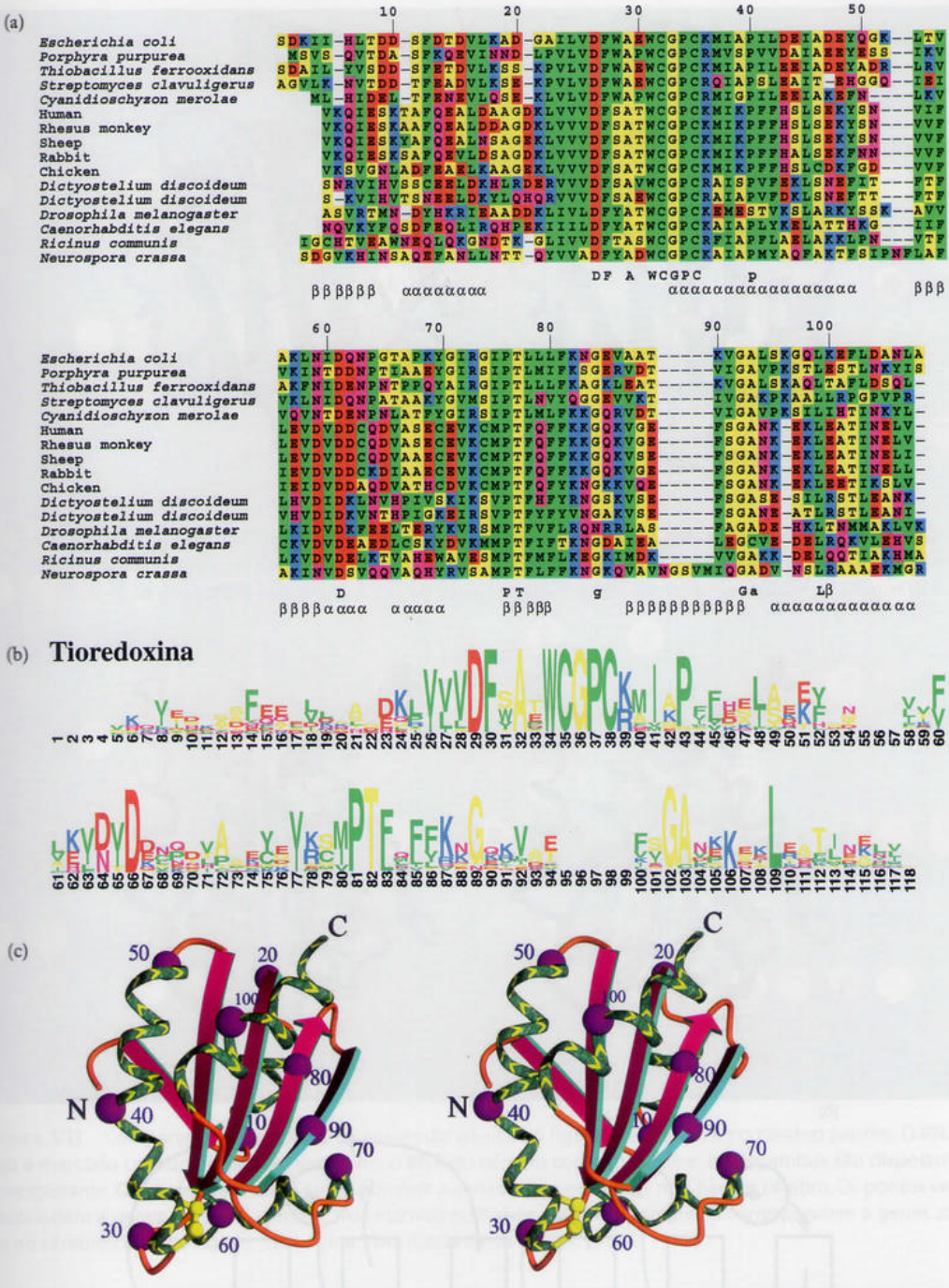


Gravura II FISH (hibridização *in situ* por fluorescência) pode detectar a presença de marcadores de locos específicos e visualizar sua posição no cromossomo. O marcador para a região centromérica do cromossomo 20, para identificar as duas cópias homólogas do cromossomo que aparecem na metáfase, é mostrado em vermelho. O marcador para D20S108 é mostrado em verde, dentro da região 20q11.2-13.1, a qual está presente em uma cópia do cromossomo 20 e ausente da outra (ver seta). Esta célula veio de um paciente sofrendo de policitemia rubra vera (um aumento anormal das células do sangue, principalmente dos eritrócitos, devido a uma anomalia na medula óssea). Acredita-se que a região removida do braço longo do cromossomo 20 contenha gene(s) supressor(es) de tumor, cuja perda contribui para o desenvolvimento de leucemias (ver página 95). (Cortesia da Dra. E. Nacheva, Department of Academic Haematology, Royal Free & University College London School of Medicine, London.)

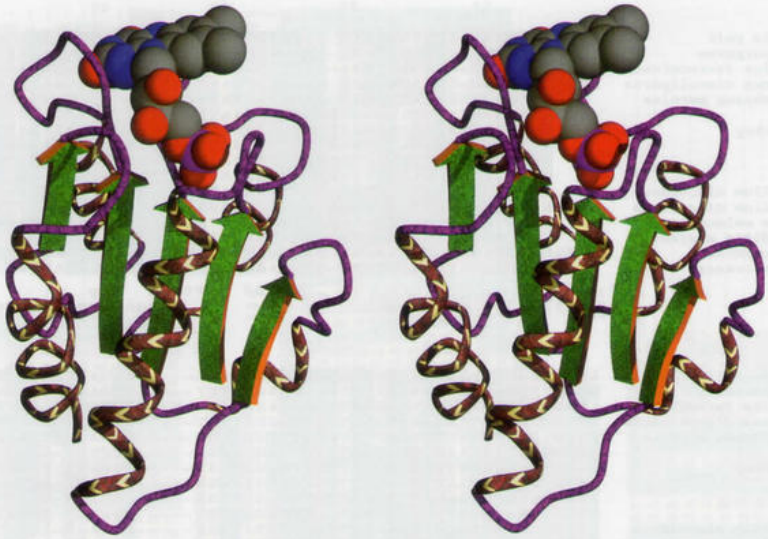
Elastases de mamíferos



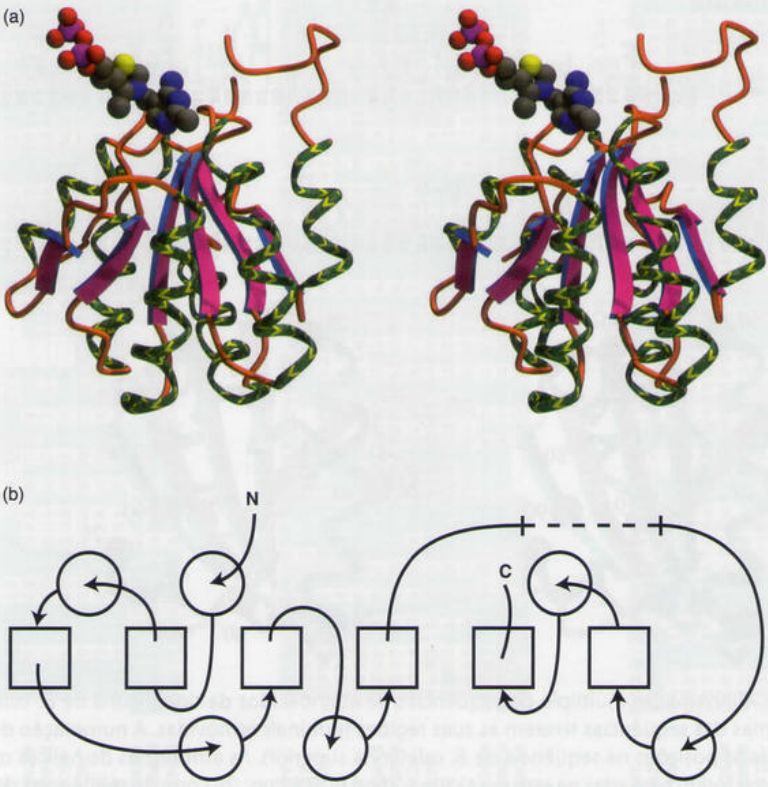
Gravura III Alinhamento múltiplo de seqüências de elastases de mamíferos (ver páginas 168-169).



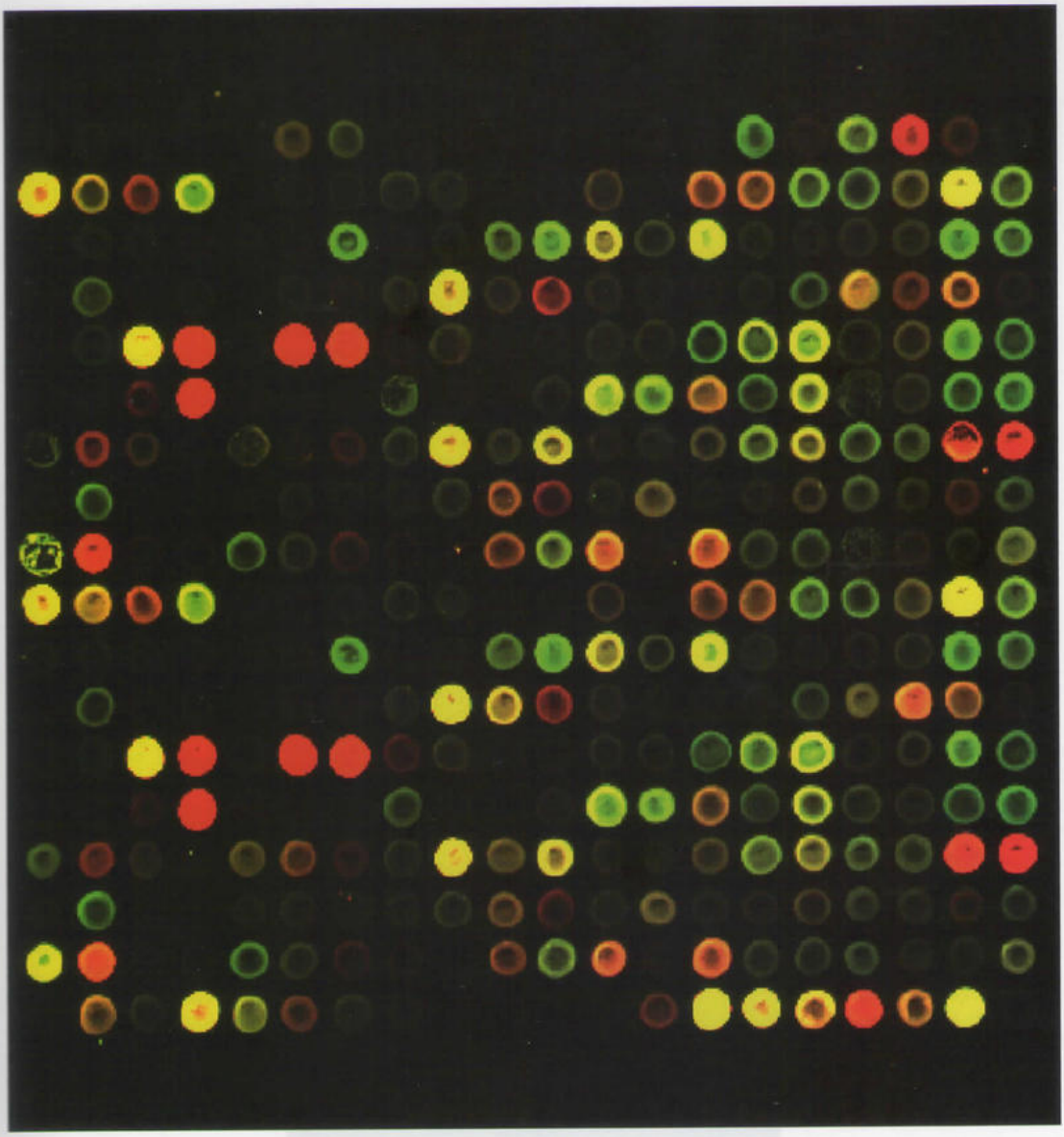
Gravura IV (a) Alinhamento múltilplo de seqüências de aminoácidos da tioredoxina de *E. coli* com seqüências homólogas. Algumas das seqüências tiveram as suas regiões terminais removidas. A numeração dos resíduos nesta tabela corresponde às posições na seqüência de *E. coli* (linha superior). As atribuições de hélices α e folhas β para a tioredoxina de *E. coli* foram baseadas na estrutura cujo código PDB é 2TRX. (b) Logo de seqüências derivado do alinhamento múltiplo de seqüências. (c) A estrutura da tioredoxina de *E. coli* [código PDB: 2TRX] contém uma folha β central com cinco fitas ladeada por hélices α. A numeração dos resíduos corresponde àquela na tabela de alinhamento múltiplo de seqüências. As regiões N- e C-terminais também se encontram destacadas. As esferas indicam as posições dos átomos Cα a cada dez resíduos. A ponte dissulfeto reativa entre Cis32 e Cis35 aparece em amarelo* (ver página 207).



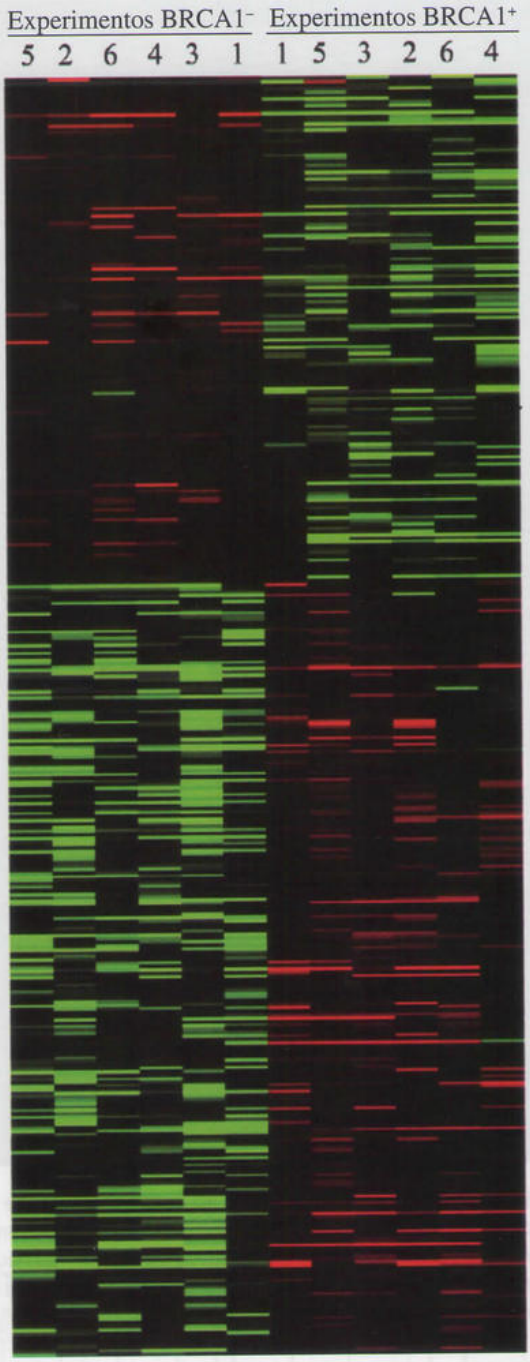
Gravura V Flavodoxina de *Clostridium beijerinckii*, ligando-se ao co-fator FMN [código PDB: 5NLL]. As setas largas representam as fitas da folha β. A caracterização desta estrutura com base na classificação hierárquica de estruturas de proteínas, baseada no banco de dados SCOP, está descrita na página 259.



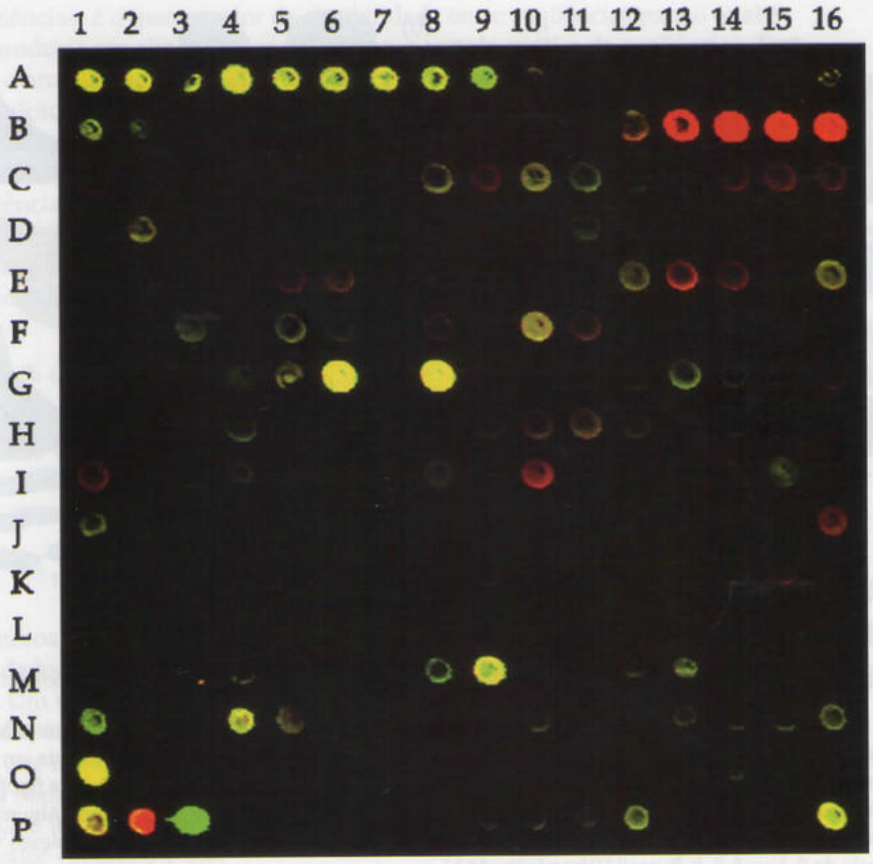
Gravura VI O domínio de ligação à tiamina da piruvato descarboxilase de levedura. Domínios de ligação à tiamina, identificados por M. Gerstein como um dos cinco padrões de enovelamento mais comuns, foram encontrados em arqueobactérias, bactérias e eucariotos. (a) Estrutura tridimensional. (b) Diagrama esquemático da topologia (ver página 285).



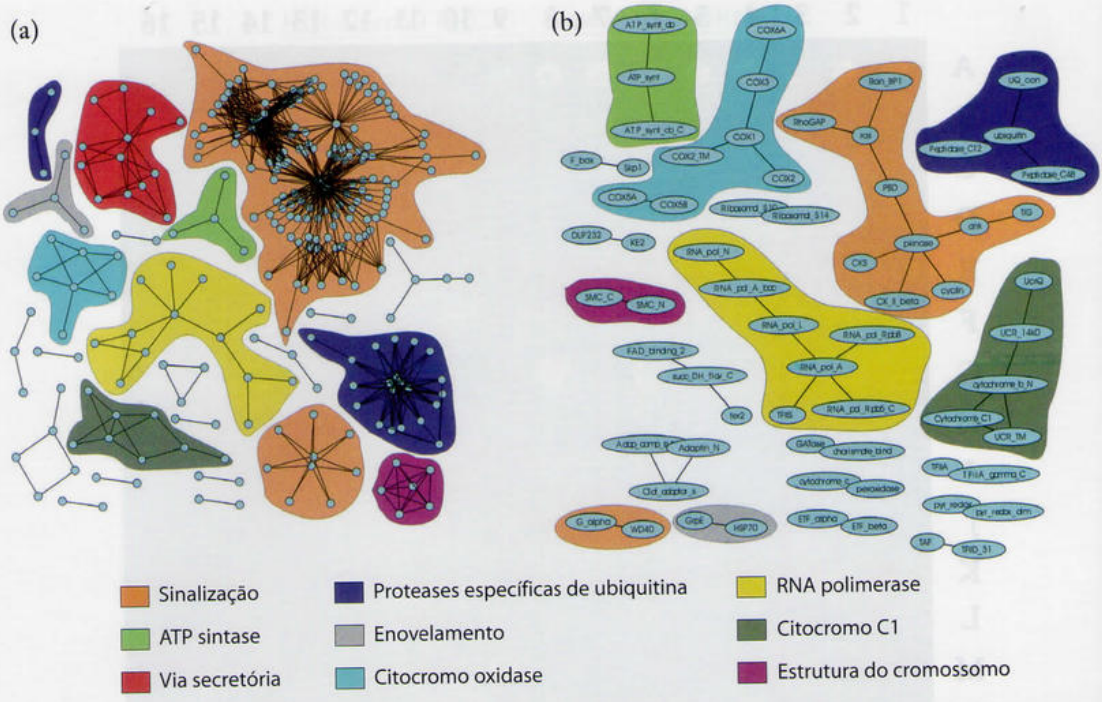
Gravura VII Comparação dos padrões de expressão gênica no fígado (vermelho) e no cérebro (verde). O RNA do fígado é marcado com um fluoróforo vermelho, o RNA do cérebro com um verde e, então, ambos são dispostos em um microarranjo. Os pontos vermelhos correspondem a genes ativos no fígado, mas não no cérebro. Os pontos verdes correspondem a genes ativos no cérebro, mas inativos no fígado. Os pontos amarelos correspondem a genes ativos tanto no cérebro como no fígado (ver página 314). (Cortesia do Dr. P. A. Lyons.)



Gravura VIII Agrupamento de dados de expressão gênica em células expressando níveis elevados e reduzidos de BRCA1. Experimentos BRCA1$^-$, com níveis reduzidos de expressão de BRCA1, aparecem nas seis colunas da esquerda. Experimentos BRCA1$^+$, com níveis elevados de expressão de BRCA1, são mostrados nas seis colunas da direita. A intensidade da cor reflete a proporção da expressão em relação a um controle. A cor vermelha revela genes com níveis de expressão mais elevados em resposta ao BRCA1. A cor verde mostra genes com níveis de expressão mais reduzidos em resposta ao BRCA1. Fonte: Welcsh, P. L., Lee, M. K., Gonzales-Hernandez, R. M., Black, D. J., Mahadevappa, M., Swisher, E. M., Warrington, J. A. & King, M.-C. (2002). BRCA1 transcriptionally regulates genes involved in breast tumourigenesis. *Proc. Natl. Acad. Sci. USA*, **99**, 7560-7566. Reproduzida com permissão (ver página 320).



Gravura IX O efeito do tratamento por quatro horas com isoniazida nos perfis de expressão de mRNA de 203 ORFs de *M. tuberculosis*. Vermelho = expresso em células tratadas com isoniazida, Verde = expresso em células sem tratamento, Amarelo = expresso tanto em células tratadas como em células sem tratamento. A série de pontos vermelhos no lado direito superior corresponde a genes do grupo de genes do sistema FAS-II. (Fonte: Wilson, M., DeRisi, J., Kristensen, H. H., Imboden, P., Rane, S., Brown, P. O. & Shoolnik, G. K. (1999). Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization, *Proc. Natl. Acad. Sci. USA*, **96**, 12833-12838. Reproduzida com permissão). (Ver páginas 328-329.)



Gravura X Parte de uma rede de interações de proteínas de levedura. A parte A (à esquerda) descreve as interações de proteínas individuais e a parte B (à direita) mostra as interações dentro de uma sub-rede, baseada em representações de diferentes famílias de proteínas, em diferentes categorias funcionais, ligadas na parte A. Esta figura é baseada em dados estruturais e modelagem. Cada relação implica uma interação física entre as proteínas. Algumas das interações envolvem complexos estáveis (por exemplo, a RNA polimerase II); outras envolvem complexos transitórios. (Figura cortesia de P. Aloy & R. B. Russell.) (Ver página 347.)

entre seqüências, é o menor valor de similaridade entre seqüências para o qual é possível produzir um alinhamento correto apenas pela análise de seqüências. É, portanto, o nível apropriado a ser escolhido para trabalhar (ver Quadro Matrizes de substituição utilizadas para a pontuação da similaridade entre seqüências de aminoácidos). Diversos autores derivaram matrizes de substituição apropriadas para uso em diferentes intervalos de similaridade global entre seqüências.

A ocorrência de inversões, tanto diretamente quanto por uma ou mais mudanças adicionais, produz uma diminuição aparente nas taxas de mutação conforme as seqüências divergem progressivamente. A relação entre o escore PAM e a % de identidade entre seqüências é:

PAM	0	30	80	110	200	250
% de identidade	100	75	50	60	25	20

A matriz PAM250 de M. O. Dayhoff é mostrada no Quadro. Ela expressa escores como valores de **probabilidades logarítmicas (*log-odds*)**:
Escore da mutação $i \leftrightarrow j$

$$= log_{10} \frac{\text{taxa de mutações i} \leftrightarrow \text{j observada}}{\text{taxa de mutação esperada das freqüências de aminoácidos}}$$

Os números são multiplicados por 10, simplesmente para evitar casas decimais. Os dados de entrada da matriz refletem as probabilidades dos eventos de mutação. Um valor de +2 – por exemplo, C ↔ S – significa que em seqüências relacionadas espera-se que esta mutação ocorra com uma freqüência 1,6 vezes maior do que mutações ao acaso. O cálculo é realizado da seguinte forma: a entrada 2 na matriz corresponde de fato ao valor 0,2 por causa do escalonamento. O valor 0,2 é $\log_{10}$ do valor relativo esperado para a mutação. Como $\log_{10}$ (1,6) = 0,2, o valor esperado é 1,6.

A probabilidade de ocorrência de dois eventos de mutação independentes é o produto das suas probabilidades. Utilizando-se logaritmos (*logs*), obtemos valores de escores que podem ser adicionados em vez de multiplicados, uma conveniência computacional.

As matrizes BLOSUM

S. Henikoff e J. G. Henikoff desenvolveram a família de matrizes BLOSUM para pontuar substituições de aminoácidos em comparação de seqüências. O objetivo era substituir a matriz Dayhoff por outra com melhor desempenho na identificação de relações distantes entre seqüências, utilizando um conjunto muito maior de dados que se tornou disponível desde o trabalho de Dayhoff.

As matrizes BLOSUM são baseadas no banco de dados BLOCKS de seqüências de proteínas alinhadas, daí o nome BLOcks SUbstitution Matrix (matriz de substituição baseada em BLOCKS). A partir de regiões de alinhamentos de proteínas estreitamente relacionadas e sem lacunas, Henikoff e Henikoff calcularam a razão do número de pares de aminoácidos observados em cada posição com o número de pares esperados pelas freqüências de ocorrência geral dos aminoácidos. Assim como a matriz de Dayhoff, os resultados são expressos como *log-odds*. Para evitar a superestimação dos escores de seqüências estreitamente relacionadas, os Henikoffs substituíram grupos de proteínas que apresentam identidade acima de um limiar por uma única seqüência representativa ou utilizaram os valores médios de escore. Um limiar de 62% origina a matriz de substituição comumente utilizada BLOSUM62 (ver Quadro). Esta matriz é oferecida

Matrizes de substituição utilizadas para a pontuação da similaridade entre seqüências de aminoácidos

Linhas e colunas estão em ordem alfabética do nome dos aminoácidos, de acordo com o código de três letras. Apenas os triângulos inferiores das matrizes são mostrados, pois as matrizes de probabilidades de substituição são simétricas. (Não porque temos certeza de que a taxa de qualquer substituição é a mesma que a taxa da substituição inversa, mas porque não podemos determinar a diferença entre as duas taxas.)

A matriz Dayhoff PAM250:

Ala (A)	2																		
Arg (R)	−2	6																	
Asn (N)	0	0	2																
Asp (D)	0	−1	2	4															
Cis (C)	−2	−4	−4	−5	12														
Gln (Q)	0	1	1	2	−5	4													
Glu (E)	0	−1	1	3	−5	2	4												
Gli (G)	1	−3	0	1	−3	−1	0	5											
His (H)	−1	2	2	1	−3	3	1	−2	6										
Ile (I)	−1	−2	−2	−2	−2	−2	−2	−3	−2	−5									
Leu (L)	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6								
Lis (K)	−1	3	1	0	−5	1	0	−2	0	−2	−3	5							
Met (M)	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6						
Phe (F)	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9					
Pro (P)	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6				
Ser (S)	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2			
Tre (T)	1	−2	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3		
Trp (W)	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17	
Tir (Y)	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10
Val (V)	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

A matriz BLOSUM62:

Ala (A)	4																		
Arg (R)	−1	5																	
Asn (N)	−2	0	6																
Asp (D)	−2	−2	1	6															
Cis (C)	0	−3	−3	−3	9														
Gln (Q)	−1	1	0	0	−3	5													
Glu (E)	−1	0	0	2	−4	2	5												
Gli (G)	0	−2	0	−1	−3	−2	−2	6											
His (H)	−2	0	1	−1	−3	0	0	−2	8										
Ile (I)	−1	−3	−3	−3	−1	−3	−3	−4	−3	4									
Leu (L)	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4								
Lis (K)	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5							
Met (M)	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5						
Phe (F)	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6					
Pro (P)	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7				
Ser (S)	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4			
Tre (T)	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5		
Trp (W)	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	
Tir (Y)	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7
Val (V)	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

por todos os programas como uma opção, sendo a escolha-padrão na maioria. As matrizes BLOSUM, e outros parâmetros de substituição derivados recentemente, substituíram a matriz de Dayhoff nas aplicações de bioinformática.

A pontuação de inserções/deleções ou a "penalização de lacunas"

Para criar um esquema de pontuação completo para alinhamentos, é necessária, além da matriz de substituição, uma forma de pontuar as lacunas. Qual é a importância de inserções e deleções em relação às substituições? Distinga uma introdução de lacuna:

```
aaagaaa
aaa-aaa
```

de uma extensão de lacuna:

```
aaaggggaaa
aaa----aaa
```

Para o alinhamento de seqüências de DNA, o programa de alinhamento CLUSTAL-W recomenda o uso de matrizes de identidade para substituições (+1 para pareamentos idênticos e 0 para desigualdades ou malpareamentos), penalidade para introdução de lacunas igual a 10 e penalidade para extensão de lacunas igual a 0,1 por resíduo. Para alinhamento de seqüências de proteínas, as recomendações são utilizar a matriz BLOSUM62 para substituições e penalidades para introdução e extensão de lacunas iguais a 11 e 1, respectivamente.

Calculando o alinhamento de duas seqüências

Agora que possuímos um esquema de escore, podemos aplicá-lo para encontrar alinhamentos ótimos – procuramos o alinhamento que maximize o escore. Um algoritmo famoso para determinar os alinhamentos ótimos globais de duas seqüências é baseado em uma técnica matemática chamada programação dinâmica (detalhes são descritos no final desta seção). Esse algoritmo tem sido extremamente importante na biologia molecular. Duas de suas características notáveis são:

- A boa notícia é que o método fornece, com certeza, o ótimo *global*. Ele irá encontrar o alinhamento de *melhor* escore, dada a escolha de parâmetros – matriz de substituição (M) e penalidades para lacunas – sem aproximações.
- A má notícia é que muitos alinhamentos podem apresentar o mesmo escore ótimo. E nenhum desses alinhamentos necessariamente corresponde ao alinhamento biologicamente correto. Por exemplo, na comparação das cadeias α e β da hemoglobina de galinha, W. Fitch e T. Smith encontraram 17 alinhamentos com o mesmo escore ótimo, dos quais um era correto (com base nas estruturas, um último recurso). Há 1.317 alinhamentos com escores a 5% do ótimo.

Outro item das más notícias é técnico: o tempo necessário para alinhar duas seqüências de comprimento n e m é proporcional a $n \times m$, pois este é o tamanho da matriz de edição que precisa ser preenchida. Isso significa que o método de programação dinâmica não é uma técnica conveniente para ser utilizada em consulta a banco de dados de seqüências para encontrar seqüências similares a uma seqüência-alvo, e ainda menos conveniente para alinhamentos "todos-contra-todos". O problema da consulta a bancos de dados é, de fato, o problema de parear (identificar homologia) uma seqüência-alvo com uma região de uma seqüência muito longa, equivalente à extensão total do banco de dados.

Variações e generalizações

Variações do método de programação dinâmica aplicadas a duas questões de alinhamentos relacionadas (ver também Quadro na página 46):

- ***alinhamento global***: encontrar o melhor alinhamento de uma seqüência inteira com outra seqüência inteira.

- ***alinhamento local***: encontrar o melhor alinhamento de algum segmento de uma seqüência contra algum segmento de outra seqüência (isso inclui testar um banco de dados contra uma única seqüência considerando o banco de dados como uma única seqüência muito longa).

O algoritmo de alinhamento global foi aplicado pela primeira vez a alinhamentos de seqüências biológicas por S. B. Needleman e C. D. Wunsch. T. Smith e M. Waterman o modificaram para identificar alinhamentos locais.

Métodos aproximados para a análise rápida de bancos de dados

É rotineiro comparar genes de um novo genoma contra bancos de dados em busca de similaridade com outras seqüências. Métodos aproximados podem detectar relações próximas com boa precisão e rapidamente, mas são inferiores aos métodos exatos na detecção de relações muito distantes. Na prática, seu desempenho é bastante satisfatório nos muitos casos em que a seqüência-alvo é bastante similar a uma ou mais seqüências do banco de dados e, portanto, é válido testá-los primeiro.

Um método aproximado típico tomaria um número inteiro pequeno k, e determinaria todas as instâncias de cada k-tupla de resíduos na seqüência-alvo que ocorrem em qualquer seqüência do banco de dados. Uma seqüência candidata é uma seqüência do banco de dados que contenha o maior número de pareamentos idênticos para as k-tuplas, com espaçamentos equivalentes na seqüência-alvo e na seqüência candidata. Para um conjunto selecionado de seqüências candidatas, são realizados cálculos de alinhamentos ótimos aproximados, com restrições que diminuem o tempo e o espaço de busca de forma que os caminhos através da matriz considerada são restritos a regiões próximas às diagonais que contenham mais pareamentos idênticos com as k-tuplas. Existem diversas variações deste tema.

O algoritmo de programação dinâmica para o alinhamento ótimo de seqüências par a par*

Uma tabela contendo implicitamente todos os alinhamentos possíveis pode ser construída como uma matriz similar à utilizada no desenho do gráfico de pontos. Os resíduos de uma seqüência indexam as linhas, e os resíduos da outra seqüência indexam as colunas. Qualquer caminho através da matriz, do canto superior esquerdo para o canto inferior direito, corresponde a um alinhamento (ver Figura 4.1). O desafio é encontrar o caminho com menor custo, e a dificuldade é que existe um grande número de caminhos a serem considerados.

Como exemplo, suponha que você queira dirigir de Malmö, no sul da Suécia, a Tromsø, no norte da Noruega (ver Figura 4.3). Sua rota consistirá em um número de fragmentos, que o levam através de uma sucessão de cidades intermediárias. Existem muitas escolhas de combinações diferentes de segmentos que produzem um caminho contínuo e completo.

O método computacional para encontrar o caminho ótimo começa com a atribuição de uma medida numérica do "custo" de cada um dos possíveis segmentos

* Esta seção é opcional. Leitores que ficarem em dúvida podem procurar mais informações em Lesk, A. M. (1988), *TATA for now...*, *Trends Biochem. Sci.* **13**, 410.

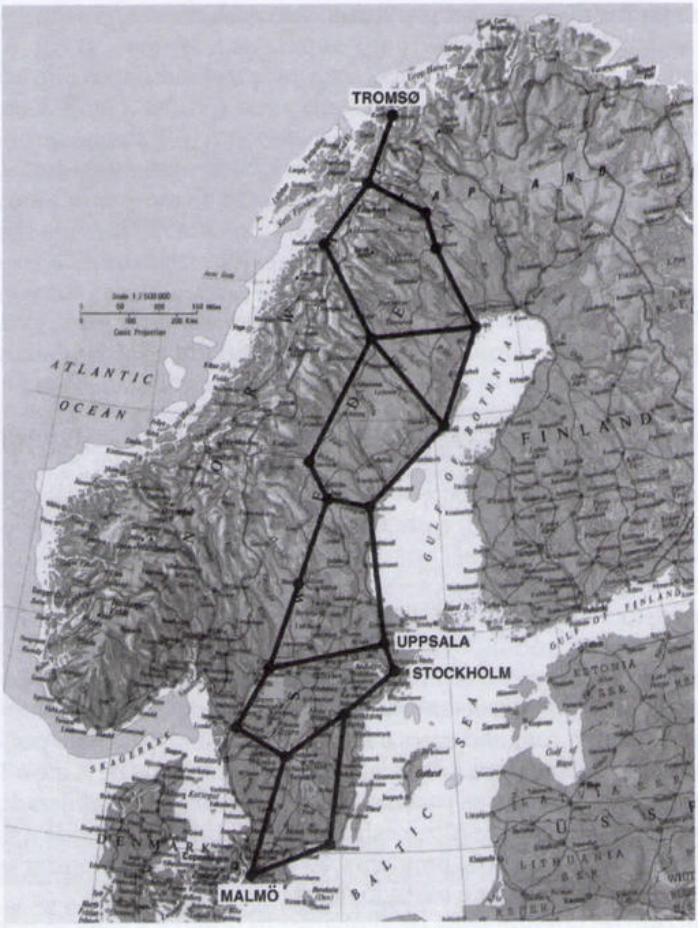
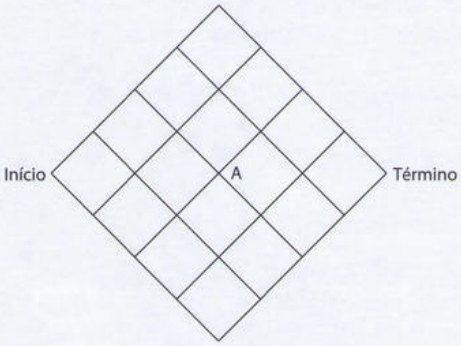


Figura 4.3 Rotas possíveis de Malmö a Tromsø. Como você pode determinar uma rota ótima? (© Collins Bartholomew 1980. Reproduzido com permissão da Editora HarperCollins.)

individuais do caminho. Esse "custo" não é simplesmente uma questão financeira, mas uma estimativa mais geral das suas preferências relativas por diferentes partes do caminho. A distância percorrida será claramente um componente importante do custo, mas outros fatores como qualidade das estradas e oportunidades de turismo também contribuem. Para qualquer rota selecionada, o custo total da viagem é a soma dos custos dos segmentos individuais. Certamente, é ineficiente percorrer repetidamente qualquer segmento, ou visitar a mesma cidade duas vezes, de modo que você concordará que qualquer ponto intermediário estará localizado ao norte do ponto anterior. Este formalismo é expresso em termos de minimização de um custo e não da maximização de um escore; para os nossos propósitos, estes dois pontos são equivalentes. Um algoritmo pode explorar as combinações possíveis para determinar a rota completa ótima.

A seguir encontra-se uma versão abstrata do problema:



Para compreender a idéia essencial da programação dinâmica, primeiro considere: Quantos caminhos entre o Início e o Término passam por A? Existem seis caminhos do Início até A. (Descreva-os.) Dessa forma, por simetria, existem seis caminhos de A até o Término, e um total de 36 caminhos entre o Início e o Término passando por A. (Por quê?) Assumindo que tenhamos atribuído custos aos passos individuais, é necessário checar todos os 36 caminhos para encontrar o caminho de custo mínimo entre o Início e o Término, passando por A? Não – aqui vai uma observação crucial: *A escolha do melhor caminho de A até o Término é independente da escolha do caminho do Início até A*. Se determinarmos o melhor dos seis caminhos do Início até A, e determinarmos o melhor dos seis caminhos de A até o Término, o melhor caminho entre o Início e o Término, passando por A, será o melhor caminho do Início até A *seguido* pelo melhor caminho de A até o Término. Não mais do que 12 caminhos passando por A precisam ser considerados.

Até simplificações maiores são possíveis pela subdivisão repetida e sistemática do problema. O método de programação dinâmica para encontrar um caminho ótimo através da matriz é baseado nesta idéia.

O enunciado de um problema de alinhamento ótimo e a solução por programação dinâmica tem a seguinte forma: dadas duas seqüências de caracteres, possivelmente de comprimentos diferentes: $A = a_1a_2...a_n$ e $B=b_1b_2...b_m$, onde cada a_i e b_j é um membro de um conjunto alfabético $\mathcal{A}$, considere seqüências de operações de edição que convertam A e B em uma seqüência comum. Operações de edição individuais incluem:

Substituição de b_j por a_i – representada por (a_i, b_j).
Deleção de a_i da seqüência A – representada por (a_i, ϕ).
Deleção de b_j da seqüência B – representada por (ϕ, b_j).

Se estendermos o conjunto alfabético para incluir o caractere nulo ϕ: $\mathcal{A}^+ = \mathcal{A} \cup \{\phi\}$, uma seqüência de operações de edição é um conjunto de pares ordenados (x, y), com x, y $\in \mathcal{A}^+$.

Uma *função de custo d* é definida em operações de edição da seguinte forma:

$d(a_i, b_j)$ = custo de uma mutação em um alinhamento no qual a posição i da seqüência A corresponde à posição j da seqüência B, e a mutação substitui $a_i \leftrightarrow b_j$.

$d(a_i, \phi)$ ou $d(\phi, b_j)$ = custo de uma deleção ou inserção.

Defina a distância mínima ponderada entre as seqüências A e B como

$$D(A, B) = \min_{A \to B} \Sigma d(x, y)$$

onde $x, y \in A^+$ e o mínimo é tomado sobre todas as seqüências de operações de edição que convertem A e B em uma seqüência comum.

O problema é encontrar $D(A, B)$ e um ou mais dos alinhamentos correspondentes a este valor.

Um algoritmo que resolva este problema em um tempo $\mathcal{O}(mn)$ cria uma matriz $\mathcal{D}(i, j)$, $i = 0, ...n$; $j=0, ...m$, tal que $\mathcal{D}(i, j)$ seja a distância mínima entre as seqüências que consistem nos primeiros i caracteres de A e nos primeiros j caracteres de B. Então, $\mathcal{D}(n, m)$ será a distância mínima necessária $D(A, B)$.

O algoritmo calcula $\mathcal{D}(i, j)$ por recursão. O valor de $\mathcal{D}(i, j)$ corresponde à conversão das subseqüências iniciais $A_i = a_1a_2...a_i$ e $B_j = b_1b_2...b_j$ em uma seqüência comum com L operações de edição S_k, $k = 1, ...L$, que pode ser considerada para ser aplicada em ordens crescentes de posição na seqüência de caracteres. Considere *desfazer* a última destas operações de edição. A seqüência de operações de edição truncada resultante, S_k, $K= 1, ... L-1$, é uma seqüência de operações de edição para converter uma subseqüência de A_i e uma subseqüência de B_j em um resultado comum. Mais importante, deve ser uma seqüência *ótima* de operações de edição para estas subseqüências, de forma que se outra seqüência S'_k fosse uma seqüência de operações de menor custo para estas subseqüências, então S'_k, seguida de S_L, seria uma seqüência de operações de menor custo do que S_k para converter A_i em B_j. Dessa forma, deveria existir um método recursivo para o cálculo de $D(i, j)$.

Reconheça a correspondência entre operações de edição individuais e os passos entre as celas adjacentes da matriz (ver Figura 4.1):

$(i-1, j-1) \rightarrow (i, j)$	corresponde à substituição $a_i \rightarrow b_j$.
$(i-1, j) \rightarrow (i, j)$	corresponde à deleção de a_i de A.
$(i, j-1) \rightarrow (i, j)$	corresponde à inserção de b_j em A na posição i.

Seqüências de operações de edição correspondem a caminhos, em uma série de passos, através da matriz

$$(i_0, j_0) = (0, 0) \rightarrow (i_1, j_1) \rightarrow ... (n, m)$$

onde $0 \leq i_{k+1} - i_k \leq 1$ (para $0 \leq k \leq n-1$), $0 \leq j_{k+1} - j_k \leq 1$ (para $0 \leq k \leq m-1$). Considerando as seqüências de operações de edição possíveis e os caminhos correspondentes através da matriz, o predecessor de uma seqüência ótima de operações de edição levando de (0, 0) para (i, j), onde $i, j > 0$, deve ser uma seqüência ótima de operações de edição que leve a uma das celas $(i-1, j)$, $(i-1, j-1)$ ou $(i, j-1)$, e, de forma correspondente, $\mathcal{D}(i, j)$ deve depender apenas dos valores de $\mathcal{D}(i-1, j)$, $\mathcal{D}(i-1, j-1)$ e $\mathcal{D}(i, j-1)$ (junto, é claro, com a parametrização especificada pela função de custo d).

O algoritmo pode ser descrito da seguinte maneira:
Compute a matriz $\mathcal{D}$ de ordem $(m+1) \times (n+1)$, aplicando:

(1) as condições de início na linha superior e na coluna da esquerda:

$$D(i,0) = \sum_{k=0}^{i} d(a_k, \phi)$$

$$D(0,j) = \sum_{k=0}^{j} d(\phi, b_k)$$

Estes valores impõem a penalidade para lacunas nos resíduos que não pareiam no início de qualquer uma das seqüências.
E então

(2) as relações de recorrência:

$$D(i,j) = min\{D(i-1,j) + d(a_i,\phi), D(i-1,j-1) + d(a_i,b_j), D(i,j-1) + d(\phi,b_j)\}$$

para $i = 1, ... n$; $j = 1, ...m$. Isso significa: considere todos os três passos possíveis de $\mathcal{D}(i, j)$:

Operação	Custo cumulativo
insira uma lacuna na seqüência A	$\mathcal{D}(i-1, j) + d(a_i, \phi)$
substitua $a_i \leftrightarrow b_j$	$\mathcal{D}(i-1, j-1) + d(a_i, b_j)$
insira uma lacuna na seqüência B	$\mathcal{D}(i, j-1) + d(\phi, b_j)$

Destas, escolha o valor mínimo. Para cada cela, registre não apenas o valor $\mathcal{D}(i, j)$, mas um ponteiro indicando (uma ou mais) da(s) cela(s) $(i-1, j)$, $(i-1, j-1)$ ou $(i, j-1)$ selecionada pela operação de minimização. Note que mais de um predecessor pode fornecer o mesmo valor.

Quando os cálculos estiverem completos, $\mathcal{D}(n, m)$ será a distância ótima $D(A, B)$. Um alinhamento correspondente à seqüência de operações de edição, registrada pelos ponteiros, pode ser obtido traçando o caminho de volta através da matriz da cela (n, m) à cela inicial $(0, 0)$. Esse alinhamento correspondendo à distância mínima $D(A, B) = \mathcal{D}(n, m)$ pode não ser único.

Exemplo 4.6 Alinhamento de seqüências par a par

Alinhe as seqüências A = `ggaatgg` e B = `atg`, de acordo com o esquema simples de escore: pareamento idêntico = 0, malpareamento = 20, inserção ou deleção = 25.

Esta é a matriz após a linha superior e a coluna mais à esquerda terem sido iniciadas (itálico), e o elemento na segunda linha e segunda coluna ter sido inserido como **20** (negrito):

	ϕ	a	t	g
ϕ	*0*	*25*	*50*	*75*
g	*25*	**20**		
g	*50*			
a	*75*			
a	*100*			
t	*125*			
g	*150*			
g	*175*			

O valor **20** foi escolhido como o mínimo de 25 + 25 (movimento horizontal, ou inserção de uma lacuna na seqüência `atg`), 0 + 20 (substituição a↔g), e 25 + 25 (movimento vertical, ou inserção de uma lacuna na seqüência `ggaatgg`). Como a substituição (movimento diagonal) forneceu o valor mínimo, a cela contendo 0, no canto superior esquerdo da matriz, é o predecessor da cela onde foi inserido o valor 20. (Se dois ou mesmo três dos movimentos possíveis apresentarem o mesmo valor, a cela resultante terá múltiplos predecessores.)

→

Esta é a matriz após a conclusão dos cálculos:

	ϕ	a	t	g
ϕ	0	← 25	← 50	← 75
	↑ ↖	↖	↖	
g	25	20	← 45	50
	↑ ↖	↑ ↖	↖	
g	50	45	40	45
	↑ ↖	↖	↑ ↖	
a	75	50	65	60
	↑ ↖	↑ ↖	↖	↑
a	100	75	70	85
	↑	↑ ↖	↖	
t	125	100	75	90
	↑	↑	↑ ↖	
g	150	125	100	75
	↑	↑	↑ ↖	↑
g	175	150	125	100

Ela inclui o caminho de retorno (*traceback*) na matriz na forma de setas apontando de cada cela para o(s) seu(s) predecessor(es). Para algumas aplicações podemos precisar apenas do valore de $D(A, B)$, mas não do alinhamento, e então não é necessário registrar os ponteiros. (Isso tem implicações no espaço necessário para realizar os cálculos.) As setas em negrito indicam o caminho do alinhamento ótimo traçando o caminho dos predecessores a partir do canto inferior direito, de volta ao canto superior esquerdo. Em alguns casos, uma cela pode apresentar dois predecessores. Eles correspondem a alinhamentos alternativos com o mesmo escore.

Existem duas celas nas quais o caminho de retorno se ramifica. Isso dá um total de quatro alinhamentos ótimos com o mesmo escore:

```
ggaatgg   ggaatgg   ggaatgg   ggaatgg
---atg-   ---at-g   --a-tg-   --a-t-g
```

Com um esquema de pesos para as lacunas, atribuindo penalidades menores para a extensão de uma lacuna do que para a sua introdução, os dois primeiros alinhamentos terão melhor escore do que os outros. No entanto, esquemas mais sofisticados de pesos para lacunas requerem fórmulas de recorrência mais complicadas para o preenchimento da matriz.

Este algoritmo determina o alinhamento *global* ótimo de duas seqüências. Ele não é apropriado para a detecção de regiões locais de alta similaridade entre duas seqüências ou para percorrer uma seqüência longa com um fragmento curto, pois ele impõe penalidades para lacunas *fora* das regiões similares. O método de T. Smith e M. Waterman resolve este problema. Sua modificação do algoritmo de programação dinâmica básico encontra alinhamentos locais ótimos, isto é, ele seleciona subseqüências das duas seqüências que sejam mais similares entre si. Suas alterações afetam:

(1) **A inicialização da matriz** – atribuição de valores na linha superior e na coluna da esquerda. No método de Smith-Waterman, a linha superior e a coluna da esquerda são indexadas como 0. Como resultado, qualquer seqüência pode ser deslocada sobre a outra antes de o alinhamento iniciar, sem a necessidade de introduzir penalidades para lacunas contra os resíduos deixados para trás.

Exemplo 4.6 (*continuação*)

(2) **Preenchimento da matriz.** No exemplo do alinhamento global, a cada passo a escolha é direcionada entre pareamento, inserção ou deleção, mesmo que nenhuma dessas escolhas seja agradável e mesmo que a sucessão de escolhas desagradáveis prejudique o escore ao longo de um caminho contendo a melhor região de identidades locais. O método de Smith-Waterman adiciona uma quarta opção: terminar a região sendo alinhada.

(3) **Escore e o caminho de retorno (*traceback*) na matriz.** O escore do alinhamento global é o número na cela da matriz localizada no canto inferior direito. No método de Smith-Waterman, para alinhamento local, ele é o valor ótimo encontrado onde quer que apareça na matriz. Para o alinhamento global, o caminho de retorno (*traceback*) para determinar o alinhamento verdadeiro começa na cela do canto inferior direito. No método de Smith-Waterman, ele começa na cela que contém o valor ótimo e retorna no sentido do canto superior esquerdo apenas pela região em que a similaridade local se estende.

O método Smith-Waterman listará apenas um alinhamento local ótimo para o nosso exemplo:

```
ggaatgg
  atg
```

Note que lacunas não aparecem fora da região pareada.
(Exemplo adaptado de: Tyler, E. C., Horton, M. R. & Krause, P. R. (1991), A review of algorithms for molecular sequence comparison, *Comp. Biomed. Res.*, **24**, 72-96.)

A significância de alinhamentos

Suponha que o alinhamento revele uma similaridade intrigante entre duas seqüências. Essa similaridade é significativa ou pode ter surgido por acaso? (Levantamos esta questão no Capítulo 1.) Para alguns fenômenos simples – jogar uma moeda ou lançar dados – é possível calcular com exatidão a distribuição esperada dos resultados e a probabilidade de qualquer resultado em particular. Para seqüências, não é trivial definir a população da qual o alinhamento é selecionado. Por exemplo, escolher seqüências aleatórias de nucleotídeos ou aminoácidos como controle ignora a predisposição natural de composições não-aleatórias.

Uma abordagem prática para este problema é descrita a seguir: se o escore do alinhamento observado não é melhor do que seria esperado de uma *permutação aleatória* da seqüência, então é provável que ele tenha ocorrido por acaso. Podemos misturar aleatoriamente uma das seqüências, muitas vezes, realinhar cada resultado com a segunda seqüência (mantida fixa) e coletar a distribuição dos escores resultantes. A Figura 4.4 mostra um resultado típico. Para consultas a bancos de dados, use a população de resultados retornados do banco de dados inteiro como a população com a qual se deve medir a estatística.

Figura 4.4 Os escores do alinhamento local ótimo para pares de seqüências de aminoácidos aleatórias de mesma extensão seguem uma distribuição de valor extremo. Note a longa cauda à direita. Isso significa que um escore com vários desvios-padrão acima da média tem maior probabilidade de ocorrer por acaso (isto é, ser *menos* significativo) do que se os escores seguissem uma distribuição normal. O gráfico mostra a função de distribuição de probabilidade. A fórmula para a função de distribuição cumulativa correspondente, ou seja, para qualquer escore x, a probabilidade de se observar um escore $\geq x$ é:

$$P(\text{Escore} \geq x) = 1 - \exp(-Ke^{-\lambda x}),$$

onde K e λ são parâmetros relacionados com a posição do máximo e a largura da distribuição.

Claramente, se as seqüências misturadas aleatoriamente pontuarem tão bem como a seqüência original, é muito improvável que o alinhamento seja significativo. Podemos medir a média e o desvio-padrão dos escores dos alinhamentos das seqüências aleatórias e nos perguntar se o escore original da seqüência é surpreendentemente alto. O **escore Z** reflete a extensão com que o resultado original se localiza bastante deslocado de uma população de resultados:

$$\text{Escore Z} = \frac{\text{escore} - \text{média}}{\text{desvio-padrão}}$$

Um escore Z igual a 0 significa que a similaridade observada não é melhor do que a média de permutações aleatórias da seqüência, e pode muito bem ter surgido por acaso. Outros valores utilizados como medidas de significância são P, a probabilidade de o alinhamento observado ter ocorrido por acaso, e, para consultas a bancos de dados, E, o número de alinhamentos tão bons quanto o observado que se esperaria ocorrer por acaso em um banco de dados do tamanho utilizado (ver Quadro na página 204).

Muitas "regras práticas" são expressas em termos da porcentagem de resíduos idênticos em um alinhamento ótimo. Se duas proteínas tiverem mais de 45% de resíduos idênticos em seu alinhamento ótimo, elas terão estruturas bastante similares e muito provavelmente terão funções comuns ou ao menos relacionadas. Se elas tiverem mais de 25% de resíduos idênticos, elas provavelmente

Como brincar com alinhamentos e não se queimar*

Alinhamentos par a par e consultas a bancos de dados geralmente apresentam similaridades tênues entre seqüências, porém tentadoras. Como podemos decidir se estamos vendo uma relação verdadeira? Dados estatísticos não podem responder questões biológicas diretamente, mas podem nos informar a probabilidade com que uma similaridade tão boa quanto a esperada pode ocorrer, apenas por acaso, entre seqüências não-relacionadas. Para tanto, queremos comparar o resultado obtido com alinhamentos das mesmas seqüências com uma grande população. Esta população "controle" deveria ser similar, em termos de características gerais, às nossas seqüências alinhadas, mas deveria conter poucas seqüências relacionadas. Somente se o alinhamento observado se destacar da população, poderemos considerá-lo como significativo.

Com qual população de seqüências devemos comparar nosso alinhamento? Para alinhamentos par a par, podemos escolher uma das duas seqüências, gerar muitas cópias misturadas com um gerador de números aleatórios e alinhar cada cópia permutada com a segunda seqüência. Para consultas a bancos de dados, todo o banco de dados serve como população de comparação.

O alinhamento da nossa seqüência com cada membro da população-controle gera um grande conjunto de escores. Como o escore do nosso alinhamento original se classifica? Diversos parâmetros estatísticos têm sido utilizados para avaliar a significância dos alinhamentos:

- O escore Z é uma medida de quão raro o nosso alinhamento original é em termos de média e desvio-padrão dos escores da população. Se o alinhamento original tiver um escore S,

$$\text{Escore Z de S} = \frac{\text{S} - \text{média}}{\text{desvio-padrão}}$$

  Um escore Z igual a 0 significa que a similaridade observada não é melhor do que a média da população-controle e pode muito bem ter ocorrido por acaso. Quanto mais alto o escore Z, maior a probabilidade de que o alinhamento observado não tenha ocorrido simplesmente por acaso. A experiência indica que escores $Z \geq 5$ são significativos.

- Muitos programas listam o valor *P*, a probabilidade de que um alinhamento seja melhor do que o acaso. A relação entre Z e *P* depende da distribuição de escores da população-controle, a qual *não* segue uma distribuição normal.

  Um guia simples para interpretar os valores *P*:

$P \leq 10^{-100}$	alinhamento perfeito
P entre 10^{-100} e 10^{-50}	alinhamento entre seqüências quase idênticas, por exemplo, alelos ou SNPs
P entre 10^{-50} e 10^{-10}	alinhamento entre seqüências estreitamente relacionadas, certamente homólogas
P entre 10^{-5} e 10^{-1}	geralmente, um parentesco distante
$P > 10^{-1}$	alinhamento provavelmente insignificante

* N. de T. *"How to play with matches and not get burned"* é um jogo de palavras no original em inglês. *"Matches"* significa tanto fósforo quanto pareamento exato de duas seqüências de caracteres.

→

- Para consultas a bancos de dados, alguns programas (inclusive PSI-BLAST) listam os valores *E*. O valor *E* de um alinhamento é o número esperado de seqüências que dão o mesmo escore Z, ou um escore melhor, se o banco de dados for alinhado contra uma seqüência aleatória. O valor *E* é obtido ao se multiplicar o valor de *P* pelo tamanho do banco de dados. Note que *E*, mas não *P*, depende do tamanho do banco de dados. Os valores de *P* estão entre 0 e 1,0. Os valores de *E* estão entre 0 e o número de seqüências do banco de dados em que a consulta foi realizada.

  Um guia simples para interpretar os valores *E*:

$E \leq 0{,}02$	seqüências provavelmente homólogas
E entre 0,02 e 1	homologia não pode ser descartada
$E > 1$	poderia se esperar um alinhamento dessa qualidade por acaso

  Estatísticas são utilizadas como um guia útil, mas não como um substituto da consideração cuidadosa dos resultados e da análise extra dos alinhamentos que parecem promissores!

terão um padrão de enovelamento similar. Por outro lado, a observação de um baixo grau de similaridade não pode excluir a homologia. R.F Doolittle definiu a região de 18 a 25% de identidade como sendo a "zona de incerteza" (*"twilight zone"*) em que a sugestão de homologia é tentadora, porém perigosa. Abaixo da zona de incerteza está uma região onde o alinhamento par a par de seqüências nos diz muito pouco. Contudo, a falta de similaridade significativa entre seqüências não elimina a similaridade entre suas estruturas.

Embora a zona de incerteza seja uma região não confiável, não estamos completamente desamparados. Para decidir se existe uma relação genuína, a "textura" do alinhamento é importante – os resíduos similares estão isolados e espalhados ao longo da seqüência? Ou existem *"icebergs"* – regiões locais de alta similaridade (outro termo de Doolittle) que podem corresponder a um sítio ativo em comum? Pode ser necessário nos basearmos em outras informações, como ligantes ou funções compartilhadas. Claro que, se as estruturas estiverem disponíveis, poderíamos examiná-las diretamente.

Alguns exemplos ilustrativos:

- A mioglobina da baleia Sperm e a leghemoglobina de tremoço (planta do gênero *Lupinus*) possuem 15% de resíduos idênticos no alinhamento ótimo. Este valor está abaixo da definição de Doolittle de zona de incerteza. Mas também sabemos que estas duas moléculas possuem estruturas tridimensionais similares, ambas contêm grupamentos heme e ambas se ligam ao oxigênio molecular. Elas são, de fato, homólogas distantemente relacionadas.
- As seqüências das metades N- e C-terminal da rodanese possuem 11% de resíduos idênticos no alinhamento ótimo. Se elas aparecessem em proteínas independentes, não seria possível concluir, a partir apenas das seqüências, que elas são relacionadas. No entanto, o fato de elas estarem em uma mesma proteína sugere que surgiram por duplicação gênica e divergência. A grande similaridade entre suas estruturas confirma os seus parentescos.

- Como uma nota de precaução, considere as proteinases quimotripsina e subtilisina. Elas têm 12% de resíduos idênticos no alinhamento ótimo. Estas enzimas têm uma função em comum e uma tríade catalítica Ser-His-Asp comum. No entanto, elas apresentam padrões de enovelamento distintos e não são relacionadas. A sua função e mecanismo comuns é um exemplo de evolução convergente. Este caso serve como um aviso contra argumentações especiais que tentam estabelecer relações entre proteínas com seqüências diferentes com base na similaridade de suas funções e mecanismos de ação!

Alinhamento múltiplo de seqüências

"Uma única seqüência de aminoácidos é tímida, um par de seqüências homólogas sussurra, muitas seqüências alinhadas gritam." Na natureza, mesmo uma única seqüência contém toda a informação necessária para determinar a estrutura de uma proteína. Como um alinhamento múltiplo de seqüências torna esta informação mais compreensível? Tabelas de alinhamentos ressaltam padrões de conservação de aminoácidos, a partir dos quais relações distantes podem ser detectadas de forma mais confiável. Ferramentas de predição de estruturas também fornecem informações mais confiáveis quando baseadas em alinhamentos múltiplos de seqüências, e não em apenas uma seqüência.

A inspeção visual de tabelas de alinhamentos múltiplos de seqüências é uma das atividades mais profícuas que um biólogo molecular pode realizar fora da bancada. Não ouse PENSAR em não montar estas tabelas com cores diferentes para aminoácidos com propriedades físico-químicas diferentes. Um esquema de cores sensato (mas não o único) pode ser o seguinte:

Cor	Tipo de resíduo	Aminoácidos
Amarelo	não polar pequeno	Gli, Ala, Ser, Tre
Verde	hidrofóbico	Cis, Val, Ile, Leu, Pro, Phe, Tir, Met, Trp
Magenta	polar	Asn, Gln, His
Vermelho	carregado negativamente	Asp, Glu
Azul	carregado positivamente	Lis, Arg

Para ser informativo, um alinhamento múltiplo de seqüências deveria conter uma distribuição de seqüências de relacionamentos próximos e distantes. Se todas as seqüências forem estreitamente relacionadas, a informação que contém será bastante redundante, e poucas inferências poderão ser feitas. Se todas as seqüências forem distantemente relacionadas, será difícil construir um alinhamento preciso (a menos que todas as estruturas estejam disponíveis) e, nestes casos, a qualidade dos resultados e as inferências que podem sugerir são questionáveis. De maneira ideal, tem-se uma gama completa de similaridades, incluindo parentes distantes ligados através de cadeias de parentescos próximos.

Estudo de Caso 4.1: Inferências estruturais a partir do alinhamento múltiplo de seqüências de tioredoxinas

Tioredoxinas são enzimas encontradas em todas as células. Elas participam de uma ampla gama de processos biológicos, incluindo proliferação celular, coagulação sangüínea, germinação de sementes, degradação de insulina, reparo de dano oxidante e regulação enzimática. O mecanismo de ação comum a estas atividades é a redução de pontes dissulfeto de proteínas.

A Gravura IV mostra o alinhamento múltiplo de seqüências de 16 tioredoxinas. A estrutura da tioredoxina de *E. coli* contém uma estrutura central composta por uma folha β formada por cinco fitas, ladeada por hélices α; essas hélices e fitas são indicadas pelos símbolos α e β. Espera-se que outras tioredoxinas compartilhem a maior parte, mas não toda a estrutura secundária da enzima de *E. coli*. A gravura também mostra um sumário do alinhamento como *logo de seqüências*, no qual as letras de tamanhos diferentes indicam proporções diferentes de aminoácidos. O logo de seqüências foi criado por T. Schneider e M. Stephens. Este exemplo foi gerado utilizando o servidor da *web* `http://weblogo.berkeley.edu`.

Características estruturais e funcionais das tioredoxinas que podemos identificar através do alinhamento múltiplo de seqüências incluem (ver Figura 4.5 e Gravura IV):

- **A região mais conservada provavelmente corresponde ao sítio ativo.** A ponte dissulfeto entre os resíduos 32 e 35 na tioredoxina de *E. coli* faz parte do motivo conservado nesta família, WCGPC[K ou R]. Outras regiões conservadas na seqüência, incluindo os resíduos PT nas posições 75-77 e GA nas posições 92-93, estão envolvidas com a ligação ao substrato.
- **Regiões ricas em inserções e deleções provavelmente correspondem a alças de superfície. Uma posição que contenha uma Gli ou Pro conservadas provavelmente corresponde a uma volta.** Voltas correlacionadas com inserções ou deleções ocorrem nas posições 9, 20, 60 e 95. A glicina conservada na posição 92 na tioredoxina de *E. coli* faz parte de uma volta.

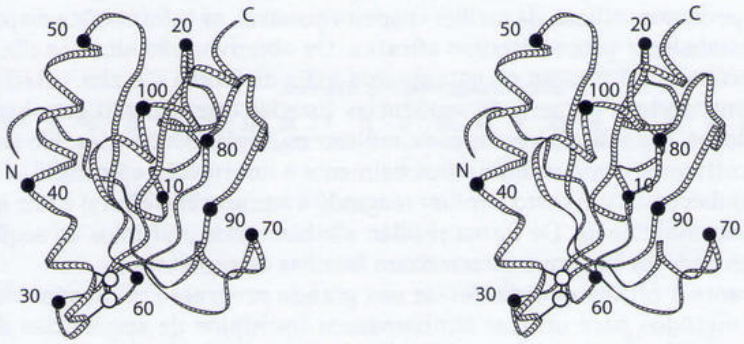


Figura 4.5 Estrutura da tioredoxina de *E. coli* [código PDB: 2TRX] (ver também Gravura IV). Os números dos resíduos correspondem aos números da tabela de alinhamento múltiplo. As porções N- e C-terminais também estão rotuladas. As esferas indicam a posição dos átomos Cα a cada dez resíduos. A ponte dissulfeto reativa entre Cis32 e Cis35 aparece entre os números 30 e 60.

→

Estudo de Caso 4.1 (*continuação*)

Esta é uma conformação pouco comum da cadeia principal, possível apenas para uma glicina (ver Capítulo 5). A prolina conservada na posição 76 da tioredoxina de *E. coli* também está associada a uma volta. Esta também é uma conformação pouco comum para a cadeia principal, acessível facilmente apenas por resíduos de prolina.

- **Um padrão conservado de hidrofobicidade, com espaçamento 2 (ou seja, um resíduo sim e um não) – com o resíduo intermediário sendo mais variável e incluindo resíduos hidrofílicos – sugere uma fita β na superfície.** Este padrão é observado na fita β entre os resíduos 50 e 60.
- **Um padrão conservado de hidrofobicidade com espaçamento de aproximadamente 4 sugere uma hélice.** Este padrão é observado na região de hélices entre os resíduos 40 e 49.

As tioredoxinas fazem parte de uma superfamília que inclui homólogas de parentesco muito mais distantes. Essas homólogas incluem a glutaredoxina (doador de hidrogênio para a redução de ribonucleotídeos na síntese de DNA), proteína dissulfeto isomerase (catalisa a troca de pontes dissulfeto malformadas no enovelamento de proteínas), fosducina (uma reguladora da via de sinalização da proteína G) e glutationa S-transferases (proteínas de defesa química). Na tabela de alinhamento múltiplo de seqüências das tioredoxinas estão implícitos padrões que deveriam ser utilizáveis na identificação destes parentes mais distantes.

Aplicações de alinhamentos múltiplos de seqüências em consultas a banco de dados

A consulta a bancos de dados por homólogas de proteínas conhecidas é um tema central da bioinformática. De fato, este tema não tolera postergações; nós o introduzimos no Capítulo 1 com um exemplo de aplicação do PSI-BLAST. Aqui reconsideramos a consulta a bancos de dados com o objetivo de tentar entender como podemos utilizar, da melhor maneira possível, as informações disponíveis para estabelecer procedimentos eficazes. Os objetivos são alta **sensibilidade** – detectando até mesmo os parentescos mais distantes – e alta **seletividade** – minimizando o número de seqüências listadas que não são genuinamente homólogas. Aqui discutiremos como utilizar múltiplas seqüências. No Capítulo 5, discutiremos como utilizar adicionalmente a informação estrutural.

Reconhecemos um rosto familiar reagindo à sua aparência geral e não a características individuais. De forma similar, alinhamentos múltiplos de seqüências contêm padrões sutis que caracterizam famílias de proteínas.

Durante a última década, fez-se um grande progresso no desenvolvimento de métodos para utilizar alinhamentos múltiplos de seqüências de proteínas conhecidas na identificação de seqüências relacionadas em consultas a bancos de dados. Os resultados são de importância fundamental para as aplicações contemporâneas da bioinformática, incluindo a interpretação de genomas. Três métodos importantes são Perfis, PSI-BLAST e Modelos Ocultos de Markov.

Perfis

Os perfis expressam os padrões inerentes a um alinhamento múltiplo de seqüências de um conjunto de seqüências homólogas. Eles possuem diversas aplicações:

- Eles permitem uma acurácia maior em alinhamentos de seqüências distantemente relacionadas.
- Conjuntos de resíduos que são altamente conservados tendem a fazer parte do sítio ativo e fornecem pistas sobre a função.
- Os padrões de conservação facilitam a identificação de outras seqüências homólogas.
- Padrões extraídos destas seqüências são úteis na classificação de subfamílias dentre um conjunto de homólogas.
- Conjuntos de resíduos que apresentam baixa conservação, e que estão sujeitos a inserções e deleções, tendem a fazer parte de alças na superfície da proteína. Esta informação tem sido aplicada no planejamento de vacinas, pois estas regiões tendem a induzir anticorpos que farão reações cruzadas significativas com as estruturas nativas.
- Muitos dos métodos de predição de estruturas são mais confiáveis se baseados em um alinhamento múltiplo de seqüências do que em um alinhamento simples. A modelagem por homologia, por exemplo, depende de forma crucial de um alinhamento de seqüências correto.

Para utilizar padrões de perfis com o objetivo de identificar homólogas, a idéia básica é parear as seqüências-alvo de um banco de dados contra as seqüências na tabela de alinhamentos, dando maior peso às posições que são conservadas em relação àquelas que são variáveis. Se uma região é absolutamente conservada, como o motivo WCGPC nas tioredoxinas, o procedimento deve insistir em encontrar este motivo. No entanto, ser tão compulsivo pode induzir ao risco de relevar parentes distantes interessantes. Portanto, alguma flexibilidade deve ser permitida.

O que é necessário é uma medida quantitativa de conservação. Para cada posição na tabela de seqüências alinhadas, considere um inventário da distribuição de aminoácidos. Por exemplo, para as posições 25-30 do alinhamento da tioredoxina:

Número do resíduo	Número de cada aminoácido																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
25	1									2								13		
26			16																	
27					16															
28																7	1		5	3
29	16																			
30			1	4									2			1	7	1		

Dada uma seqüência-alvo representando uma potencial homóloga da tioredoxina, queremos avaliar a sua similaridade com a seqüência-alvo, de forma que a concordância com as seqüências conhecidas nas posições absolutamente conservadas – por exemplo 26, 27 e 29 – contribua com um escore alto e a discordância nestas posições contribua com um escore baixo. Para posições mo-

deradamente conservadas, tais como 28, desejamos uma contribuição positiva modesta para o escore se a seqüência-alvo apresentar S ou W nesta posição, e uma contribuição ainda menor se for T ou Y. A idéia geral é pontuar cada resíduo da seqüência-alvo com base na distribuição de aminoácidos naquela posição da tabela de alinhamento múltiplo de seqüências.

Uma técnica tentadora, porém muito simplista, seria utilizar os inventários diretamente como escores. Por exemplo, se os resíduos em uma seqüência-alvo que correspondem às posições 25-30 na tioredoxina contêm a seqüência VDFSAE, este fragmento teria um escore 13 + 16 + 16 + 7 + 16 + 4 = 72. Este é quase o maior escore possível. A seqüência-alvo alternativa ACGWAP teria escore 1 + 0 + 0 + 5 + 16 + 2 = 24, um valor muito mais baixo. É claro que, para cada seqüência-alvo, teríamos de testar todos os alinhamentos possíveis com a tabela de alinhamento múltiplo e escolher o maior escore total. As seqüências com escores maiores se adaptam melhor ao padrão implícito na tabela.

Este método simples funcionaria se nossa tabela contivesse uma amostra abrangente e não tendenciosa de seqüências de tioredoxinas. Mas, apenas neste caso, um inventário simples forneceria uma representação correta das distribuições *potenciais* de resíduos em cada posição. Se nossa amostra fosse pequena, o padrão derivado provavelmente não refletiria o repertório completo. Ou, se a amostra contivesse um grande subconjunto de seqüências similares, essas seriam super-representadas nos inventários. Por exemplo, podemos ver na Gravura IV que as tioredoxinas de vertebrados formam um conjunto muito estreitamente relacionado. Se incluíssemos mais vinte seqüências de tioredoxinas de vertebrados no alinhamento, o perfil reconheceria apenas tioredoxinas de vertebrados de forma eficiente.

Matrizes de substituição sugerem como tornar o inventário "elusivo" e, portanto, mais geral.

A distribuição observada de aminoácidos em qualquer posição é uma tabela de 20 membros: $(a_1, a_2, a_3, \ldots a_{20})$, onde a_i é o número de aminoácidos do tipo i observado em uma dada posição. (Para a posição 25 das tioredoxinas, $a_1 = 1$, pois uma alanina é observada, e $a_{18} = 13$, representado as valinas.) Assim, no esquema mais simples, o escore de uma alanina na posição 25 é apenas 1; o escore de uma valina é 13; em geral, o escore de um aminoácido do tipo i é a_i. Neste esquema, as próprias linhas do inventário fornecem as tabelas a necessárias para a pontuação de cada posição.

Um esquema de escores melhor avaliaria qualquer aminoácido de acordo com a chance de ele ser substituído por um dos aminoácidos observados. Se $D(i, j)$ é a matriz de substituição de aminoácidos – talvez BLOSUM62 – então, o aminoácido i poderia ter um escore $a_1D(i, 1) + a_2D(i, 2) \ldots a_{20}D(i, 20)$. Este esquema distribui o escore entre os aminoácidos observados, com pesos de acordo com a probabilidade de substituição. Um aminoácido na seqüência-alvo poderia ter um escore alto *tanto* se ele for observado com freqüência em uma dada posição, *quanto* se houver uma alta probabilidade que ele surja por mutação dos tipos de resíduos mais comuns nesta posição. Este método é mais eficaz na detecção de parentes distantes de um conjunto limitado de seqüências conhecidas. Neste caso, o vetor de pontuação para os aminoácidos é o produto da matriz de substituição com as linhas da tabela do inventário. Um método ainda melhor é utilizar como distribuição de aminoácidos uma combinação do inventário observado e um nível geral de composição de aminoácidos.

O resultado é um conjunto de escores de probabilidades para cada aminoácido (ou lacuna) em cada posição do alinhamento, chamada de **matriz de pon-**

tuação dependente da posição (*position-specific scoring matrix*). Um método alternativo para derivação de matrizes de pontuação dependente da posição, baseado em estruturas tridimensionais, é descrito no Capítulo 5.

Dada uma seqüência-alvo, e uma matriz de pontuação dependente da posição derivada de um perfil, os cálculos necessários para encontrar um escore ótimo dentre todos os alinhamentos da seqüência-alvo com o perfil são extensões do método de programação dinâmica de alinhamento de duas seqüências.

O ponto fraco de perfis simples é que alinhamentos múltiplos de seqüências devem ser fornecidos previamente e considerados como fixos. Os métodos PSI-BLAST e Modelos Ocultos de Markov se tornam mais robustos por integrarem a etapa de alinhamento com a coleta de estatísticas.

PSI-BLAST

PSI-BLAST é um programa que realiza consultas em um banco de dados por seqüências similares a uma seqüência-alvo. Ele é um desenvolvimento do programa anterior BLAST (Basic Local Alignment Search Tool). O programa BLAST e suas variáveis (ver Quadro) verificam cada entrada no banco de dados de forma *independente* contra a seqüência-alvo. O PSI-BLAST inicia com consulta do tipo uma-de-cada-vez. Ele então deriva padrões de informação do alinhamento múltiplo de seqüências a partir das seqüências encontradas inicialmente e consulta novamente o banco de dados utilizando este padrão. PSI-BLAST então repete o processo, refinando o padrão em ciclos sucessivos.

Os programas BLAST estão disponíveis em diversos "sabores"

Programa	Tipo de seqüência-alvo	Consulta em banco de dados de
BLASTP	seqüências de aminoácidos	seqüências de proteínas
BLASTX	seqüências de nucleotídeos traduzidas	seqüências de proteínas
TBLASTN	seqüências de aminoácidos	seqüências de nucleotídeos traduzidas
TBLASTX	seqüências de nucleotídeos traduzidas	seqüências de nucleotídeos traduzidas
PSI-BLAST	seqüências de aminoácidos	bancos de dados de seqüências de proteínas

Estes programas comparam seqüências de aminoácidos com seqüências de aminoácidos, usando a matriz BLOSUM62 como padrão. Consultas envolvendo seqüências de nucleotídeos, tanto como seqüências-alvo ou como banco de dados, são realizadas com a sua tradução da seqüência de nucleotídeos para a seqüência de aminoácidos em todas as seis fases de leitura possíveis. Outro programa desta família, BLASTN, compara seqüências-alvo de nucleotídeos diretamente com bancos de dados de nucleotídeos.

O BLAST foi originalmente planejado para resolver o problema dos métodos de programação dinâmica exaustivos que são muito lentos para realizar consultas completas em grandes bancos de dados. Com freqüência, o banco de dados contém seqüências similares à seqüência-alvo. Programas menos sensíveis, porém mais rápidos, são bastante capazes de identificar essas seqüências e, se for este o

resultado desejado, ótimo. Por exemplo, se você desejar procurar por homólogas de proteínas de camundongos no genoma humano, a similaridade será alta, e um método aproximado irá encontrar essas homólogas. Mas se você desejar procurar por homólogas de uma proteína humana no genoma de *C. elegans* ou de levedura, a relação poderá ser mais sutil, e métodos mais sofisticados, porém mais lentos, podem ser necessários. (Pode ser uma surpresa, mas o tempo computacional necessário ainda deve ser levado em consideração. Embora o tempo de computação esteja deixando de ser um problema, o tamanho dos bancos de dados e o número de consultas desejadas, em todo o mundo, estão crescendo. O efeito geral é que a pressão sobre os recursos computacionais está aumentando.)

O método utilizado pelo BLAST retoma, de certo modo, o método do gráfico de pontos, verificando regiões de identidade local. Para cada entrada em um banco de dados, ele busca por pequenas regiões contíguas que pareiam com pequenas regiões contíguas da seqüência-alvo, utilizando uma matriz de substituição, mas sem permitir a introdução de lacunas. Um método no qual regiões candidatas de *extensão fixa* são identificadas inicialmente pode se tornar bastante rápido com o uso de tabelas de consulta.

Uma vez que o BLAST identifica uma região bem pareada, ele tenta estendê-la. Em algumas versões, a inserção de lacunas é permitida. A saída do BLAST é um conjunto de segmentos com pareamentos locais. No exemplo do Capítulo 1:

```
My.care.is.loss.of.care,.by.old.care.done,
   |||||||||     ||||||||||||     ||||||  ||
Your.care.is.gain.of.care,.by.new.care.won
```

mesmo um algoritmo bastante simples poderia encontrar regiões com pareamento de quatro resíduos contíguos, e então combiná-las e estendê-las (ver Problema 4.5).

O PSI-BLAST, utilizando buscas iterativas de padrões (ver Quadro), é muito mais poderoso para identificar relações distantes do que o alinhamento simples, par a par, do BLAST. O PSI-BLAST identifica corretamente três vezes mais homólogas do que o BLAST nas regiões onde a identidade entre as seqüências é inferior a 30%. PSI-BLAST é, portanto, um método muito útil para a análise de genomas completos.

Os únicos métodos baseados inteiramente na análise de seqüência que apresentam melhores resultados que o PSI-BLAST são os Modelos Ocultos de Markov. Esses modelos são descritos na próxima seção. Para obter um desempenho melhor, é necessário utilizar informações estruturais explicitamente. Isso será discutido no próximo capítulo.

O fluxo de trabalho do PSI-BLAST

1. Testa cada seqüência do banco de dados escolhido por regiões de similaridade local com a seqüência-alvo de forma independente, utilizando uma busca do tipo BLAST, mas permitindo a introdução de lacunas.
2. Armazena os resultados significativos. Constrói uma tabela de alinhamento múltiplo de seqüências entre a seqüência-alvo e os pareamentos locais significativos.
3. Constrói um perfil a partir do alinhamento múltiplo de seqüências.
4. Consulta o banco de dados novamente com o perfil, ainda procurando apenas por pareamentos locais.

→

5. Decide quais seqüências identificadas são estatisticamente significativas e mantém apenas estas.
6. Retorna ao passo 2, até que um ciclo não produza mais alterações. Isso corresponde à parte "Iterativa" no título do programa (PSI, Position Specific Iterated).

Modelos ocultos de Markov

Um modelo oculto de Markov (ou HMM, do inglês Hidden Markov Model) é uma estrutura computacional para descrever padrões sutis que definem famílias de seqüências homólogas. HMMs são ferramentas poderosas para a detecção de parentescos distantes e para a predição de padrões de enovelamento de proteínas. Este é o único método baseado inteiramente em seqüências – ou seja, sem utilizar explicitamente informações estruturais – que é competitivo com o PSI-BLAST na identificação de homólogos distantes. O método também tem bom desempenho no reconhecimento de enovelamentos, como demonstrado nos projetos CASP.

Em um HMM existem alinhamentos múltiplos de seqüências. No entanto, HMMs geralmente são apresentados como *procedimentos para a geração de seqüências*. Uma tabela convencional de alinhamento múltiplo de seqüências pode também ser utilizada para gerar seqüências, pela seleção de aminoácidos em posições sucessivas, cada aminoácido sendo escolhido a partir de uma distribuição de probabilidades dependente da posição e derivada de um perfil. Mas HMMs são mais gerais do que perfis:

1. Eles incluem a possibilidade de introduzir lacunas na seqüência gerada, com penalidades para lacunas dependentes da posição.
2. A aplicação de perfis exige que o alinhamento múltiplo de seqüências seja especificado inicialmente: a estatística do padrão é então derivada do alinhamento. HMMs realizam o alinhamento e a atribuição de probabilidades ao mesmo tempo.

A estrutura interna de um HMM mostra o mecanismo para a geração de seqüências (Figura 4.6). Comece em Início, e siga alguma cadeia de setas até chegar ao Fim. Cada seta o leva a um estado do sistema. Em cada estado (1) você realiza alguma ação – gera um resíduo, talvez – e (2) escolhe uma seta que o leva ao próximo estado. A ação e a escolha do estado sucessor são governadas por conjuntos de probabilidades. A cada estado que gera um resíduo estão associadas uma distribuição de probabilidade para os 20 aminoácidos e uma segunda distribuição de probabilidade para a escolha do estado sucessor. Essas duas distribuições de probabilidades são calibradas para codificar informações sobre uma família de seqüências em particular. Dessa forma, a mesma estrutura geral pode ser especializada para diferentes famílias de seqüências.

A dinâmica do sistema é tal que apenas o estado atual influencia a escolha do seu sucessor – o sistema não possui "memória" de sua história. Isso é característico de processos de Markov de primeira ordem, estudados no século XIX pelo matemático russo A. A. Markov. Ele distingue a sucessão de estados da sucessão de aminoácidos gerados para formar a seqüência de saída. Vários caminhos diferentes através do sistema podem gerar a mesma seqüência. Apenas a sucessão de caracteres gerada é visível; a seqüência de estados que gerou os caracteres permanece interna ao sistema, ou seja, oculta (do inglês *hidden*). Pelas distribuições de probabilidade associadas com os estados individuais, o

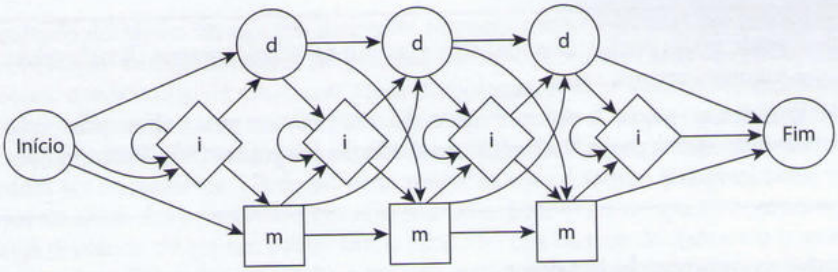


Figura 4.6 A estrutura de um Modelo Oculto de Markov (HMM). Correspondendo a cada posição de um resíduo em um alinhamento múltiplo de seqüências, o HMM contém um estado de pareamento (m) e um estado de deleção (d). Estados de inserção (i) aparecem entre as posições dos resíduos e também no início e no fim.

- Estados de pareamento geram um resíduo. Aqui, o termo pareamento significa apenas que há *algum* aminoácido tanto no modelo HMM quanto na seqüência gerada, e não necessariamente que estes aminoácidos são o *mesmo*. A probabilidade de gerar cada um dos 20 aminoácidos em cada um dos estados de pareamento é uma propriedade do modelo. Assim como nos perfis, as probabilidades são dependentes da posição.
- Estados de deleção "pulam" uma coluna no alinhamento múltiplo de seqüências. Chegar a um estado de deleção a partir de um estado de pareamento ou estado de inserção corresponde à introdução de uma lacuna, e as probabilidades desta transição refletem uma penalidade para introdução de lacuna dependente da posição. Chegar a um estado de deleção a partir de um estado de deleção anterior corresponde à extensão da lacuna.
- Estados de inserção aparecem entre duas posições sucessivas no alinhamento. Se o sistema adicionar um estado de inserção, um novo resíduo que não corresponde a uma posição da tabela de alinhamento surge na seqüência gerada. Um estado de inserção pode ser sucedido por ele mesmo, para inserir mais do que um resíduo. A sucessão de resíduos gerada pelos estados de pareamento e de inserção produz a seqüência de saída.

Após realizar a ação apropriada a cada tipo de estado (m, d ou i), outra distribuição de probabilidade governa a escolha do próximo estado. Em cada sucessão de estados possível, cada coluna do alinhamento inerente ao HMM deve ser averiguada, e deve incluir um pareamento ou uma deleção – não há nenhuma maneira de percorrer a rede sem passar através de um estado m (de pareamento) ou d (de deleção) para cada posição.

sistema captura – ou modela – os padrões intrínsecos (ou ocultos) a uma família de seqüências. Daí o nome, modelos ocultos de Markov.

Programas que aplicam HMMs na análise de seqüências biológicas podem realizar:

1. **Treinamento** Dado um conjunto de seqüências homólogas não-alinhadas, ele pode alinhá-las e ajustar as probabilidades de transição e saída de resíduos para definir um HMM que capture padrões intrínsecos às seqüências submetidas.
2. **Detecção de homólogas distantes** Dado um HMM e uma seqüência teste, calcule a probabilidade com que o HMM geraria a seqüência teste. Se um HMM treinado com uma família de seqüências conhecidas gera a seqüência teste com probabilidade relativamente alta, a seqüência teste provavelmente pertence a esta família.
3. **Alinhamento de seqüências adicionais** A probabilidade de qualquer seqüência de estados pode ser calculada a partir das probabilidades de transição individuais de estado para estado. Encontrar a seqüência de estados mais provável que o HMM utilizaria para gerar uma ou mais seqüências teste revela o seu alinhamento ótimo com a família.

Estudo de Caso 4.2: Predição de proteínas transmembrana e peptídeos sinal por Modelos Ocultos de Markov

As membranas são o invólucro das células e dos compartimentos subcelulares e organelas. As membranas contêm bicamadas fosfolipídicas. As caudas dos fosfolipídeos apontam para o interior da membrana, criando um meio orgânico de cerca de 3 nm de espessura. Os grupamentos das cabeças polares apontam para fora. Entre a região orgânica e o meio externo aquoso existem camadas de transição ricas em cadeias laterais aromáticas, de aproximadamente 1,5 nm de espessura, nas duas faces da membrana.

Muitas proteínas são planejadas para ficarem no interior das membranas. Entre as suas adaptações estão regiões que contêm principalmente resíduos apolares, que interagem com a camada orgânica da membrana. Proteínas de membrana medeiam a troca de matéria, energia e informação entre o interior celular e o seu entorno. Exemplos de funções de proteínas de membrana incluem a transdução de energia, por meio da geração ou liberação de gradientes de concentrações através de membranas da célula ou organela celular, além de recepção e transmissão de sinal.

Estima-se que, no genoma humano, cerca de 30% dos genes codifiquem proteínas de membrana. Posto que as proteínas de membrana sejam tão comuns, é importante ter ferramentas confiáveis para a sua anotação. Relativamente poucas estruturas de proteínas de membrana foram resolvidas experimentalmente. Isso coloca uma responsabilidade ainda maior nas ferramentas computacionais de análises de seqüências para identificar e caracterizar proteínas de membranas.

Aproximadamente 70% dos alvos conhecidos de drogas são proteínas transmembrana.

Muitas proteínas de membrana contêm um conjunto de sete hélices α consecutivas que atravessam a membrana, orientadas de forma aproximadamente perpendicular ao plano da membrana (ver Figura 4.7). Essas hélices são conectadas por alças que se estendem em direção ao meio aquoso (uma segunda classe de proteínas de membrana contém um barril β). Hélices transmembrana possuem tipicamente entre 15 e 30 resíduos. Apesar de serem ricas em resíduos hidrofóbicos – adequados ao ambiente apolar – elas contêm algumas cadeias laterais polares, geralmente na interface das hélices empacotadas umas contra as outras na estrutura.

Uma dica simples quanto à orientação das hélices através da membrana é a "regra do interior positivo". As alças entre as hélices se localizam inteiramente no interior ou inteiramente no exterior da célula. As alças localizadas no interior apresentam uma preponderância de resíduos carregados positivamente.

Um método simples para a predição de proteínas de membrana envolve a procura por segmentos de aminoácidos de 15-30 resíduos de extensão que sejam ricos em resíduos hidrofóbicos. No entanto, peptídeos sinal também contêm hélices hidrofóbicas: a seqüência sinal tipicamente contém uma região N-terminal carregada positivamente, seguida por uma região H helicoidal hidrofóbica, seguida por uma região C-terminal polar. Métodos para o reconhecimento de hélices transmembrana em seqüências de aminoácidos tendem a identificar as regiões H de peptídeos sinal como falso-positivos. Métodos para o reconhecimento de peptídeos sinal em seqüências de aminoácidos tendem a identificar hélices transmembrana como falso-positivos.

→

Estudo de Caso 4.2 (*continuação*)

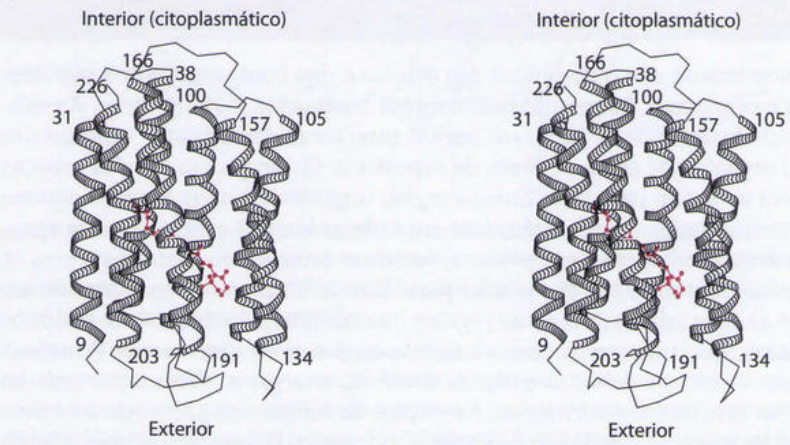


Figura 4.7 Bacteriorrodopsina da bactéria *Halobacterium salinarum* (antigamente denominada *Halobacterium halobium*) [código PDB: 2BRD], vista no plano da membrana. O ligante mostrado em vermelho é o cromóforo retinal.

Käll, Krogh e Sonnhammer treinaram modelos ocultos de Markov (HMMs) para testar simultaneamente hélices transmembrana e peptídeos sinal.* Os objetivos são os de identificar ambos ao mesmo tempo, discriminá-los nos resultados e predizer não apenas a posição das hélices transmembrana, mas as localizações – citoplasmática ou externa – das alças. O método, chamado **Phobius**, encontra-se disponível em `http://phobius.cgb.ki.se/`.

Desenvolvido a partir de trabalhos anteriores, o HMM para a predição combinada contém HMMs distintos para a predição de hélices transmembrana e para a predição de peptídeos sinal conectados em paralelo. Uma conexão une os modelos: uma transição é possível a partir do último estado do modelo de peptídeo sinal para o estado de alça voltada para o lado citoplasmático do modelo da proteína transmembrana helicoidal. Isso torna possível lidar com seqüências de proteínas que contêm uma seqüência peptídeo sinal *seguida* por uma estrutura de hélice transmembrana. Como o peptídeo sinal deve *preceder* uma estrutura helicoidal transmembrana na seqüência, nenhuma conexão *do* modelo de estrutura helicoidal transmembrana *para* o modelo de peptídeo sinal é necessária.

Phobius foi treinado com um conjunto escolhido de várias centenas de seqüências, mantendo apenas um exemplar de cada conjunto de seqüências estreitamente relacionadas. O conjunto inicial incluía proteínas de membrana e seqüências controle que não continham peptídeos sinal ou hélices transmembrana. A anotação das seqüências das proteínas transmembrana do conjunto de treinamento classificou cada resíduo como:

* Käll, L., Krogh, A. & Sonnhammer, E. L. (2004), A combined transmembrane topology and signal peptide prediction method, *J. Mol. Biol.*, **338**, 1027-1036.

- alça citoplasmática
- hélice transmembrana
- alça não-citoplasmática
- alça longa não-citoplasmática
- sítio de clivagem de peptídeo sinal
- região N-terminal de peptídeo sinal
- região H de peptídeo sinal
- região C-terminal de peptídeo sinal

As seqüências disponíveis foram separadas em dez grupos. Em dez cálculos separados, o sistema foi treinado com a junção de nove dos dez grupos, com o décimo reservado para teste. Após comparações e ajustes, um estágio final de refinamento do modelo utilizou todas as seqüências agrupadas.

A aplicação do Phobius a seqüências da bacteriorrodopsina, a proteína mostrada na Figura 4.7, fornece os seguintes resultados:

```
ID   2brd
FT   DOMAIN       7   11      NON CYTOPLASMIC.
FT   TRANSMEM    12   29
FT   DOMAIN      30   40      CYTOPLASMIC.
FT   TRANSMEM    41   63
FT   DOMAIN      64   82      NON CYTOPLASMIC.
FT   TRANSMEM    83  101
FT   DOMAIN     102  107      CYTOPLASMIC.
FT   TRANSMEM   108  129
FT   DOMAIN     130  134      NON CYTOPLASMIC.
FT   TRANSMEM   135  156
FT   DOMAIN     157  176      CYTOPLASMIC.
FT   TRANSMEM   177  199
FT   DOMAIN     200  204      NON CYTOPLASMIC.
FT   TRANSMEM   205  224
FT   DOMAIN     225  227      CYTOPLASMIC.
```

Os leitores podem comparar esta predição com a estrutura experimental (ver Problema 4.9).

O Phobius é o algoritmo mais bem-sucedido atualmente disponível para o reconhecimento de peptídeos sinal e proteínas transmembrana helicoidais e para predizer a orientação de segmentos transmembrana. Phobius é capaz de distinguir domínios H de peptídeos sinal de hélices transmembrana. O número de classificações falsas de peptídeos sinal é de 3,9%, e o número de classificações falsas de hélices transmembrana é de 7,7%. Estes resultados representam um grande avanço em relação aos métodos anteriores. É interessante notar que abordar os dois problemas ao mesmo tempo se mostrou mais eficaz do que tratá-los separadamente.

SÍTIO DA WEB ASSOCIADO

Recursos na *web*: Modelos Ocultos de Markov

Dois grupos de pesquisa especializados em aplicações biológicas de Modelos Ocultos de Markov mantêm servidores na *web* e distribuem seus programas:

R. Hughey, K. Karplus and D. Haussler (University of California at Santa Cruz, CA, USA):
`http://www.soe.ucsc.edu/research/compbio/sam.html`
`http://www.soe.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html`

S. Eddy (HHMI Janelia Farm Research Campus, Ashburn, VA, USA):
`http://hmmer.janelia.org/`

Resultados da análise de seqüências e estruturas conhecidas também estão disponíveis na *web*:

Pfam é um banco de dados de alinhamentos múltiplos de seqüências e HMMs para muitos domínios protéicos desenvolvido por A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe e E. L. Sonnhammer:
`http://www.sanger.ac.uk/Software/Pfam/`

J. Gough, K. Karplus, R. Hughey e C. Chotia geraram HMMs para todas as superfamílias do PDB:
`http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/`

Filogenia

Já vimos diversos exemplos de evolução em proteínas e genomas. Eles representam a extensão, para o nível molecular, de assuntos importantes que têm ocupado biólogos desde Darwin, e até mesmo antes. O princípio básico é que *a origem da similaridade é a ancestralidade comum*. Apesar de haver muitas exceções, originadas por evolução convergente ou transferência horizontal de genes, a importância deste princípio tanto para racionalizar observações contemporâneas quanto para estabelecer uma janela na história da vida não pode ser subestimada.

O campo da filogenia tem como objetivos trabalhar com as relações entre espécies, populações, indivíduos ou genes (o termo geral é "taxa"). Os taxa *observáveis* – por exemplo, as espécies existentes para as quais queremos elaborar o padrão de ancestralidade são chamados de "unidades taxonômicas operacionais", cuja abreviatura é OTU (do inglês Operational Taxonomic Units). A relação é considerada no sentido literal de parentesco ou genealogia, isto é, a atribuição de um esquema de descendentes de um ancestral comum (ver Quadro). Relações evolucionárias nos dão uma idéia do desenvolvimento histórico da vida (ver Quadro A escala de tempo da história da Terra). Embora as próprias moléculas não possam ser datadas, os eventos evolucionários observados no nível molecular podem ser calibrados com o registro fóssil.

Conceitos relacionados com classificação biológica e filogenia (ver página 49)

Homologia significa, especificamente, descender de um ancestral comum. Similaridade é a medida de semelhança ou diferença, independente da origem desta semelhança. A similaridade é observada em dados coletados *agora* e não envolve hipóteses históricas. Em contraste, atribuições de homologia requerem inferências sobre eventos históricos que são quase sempre inobserváveis.

→

Agrupamento é a união de itens similares, distinguindo classes de objetos que são mais similares entre si do que com os demais objetos de outras classes. A maioria das pessoas concordaria com graus de similaridade, mas o agrupamento é mais subjetivo. Na classificação de objetos, algumas pessoas preferem grandes classes, tolerando variações maiores; outras preferem classes menores, mais compactas. Estas pessoas são chamadas de *classificadoras* ou *separadoras*.

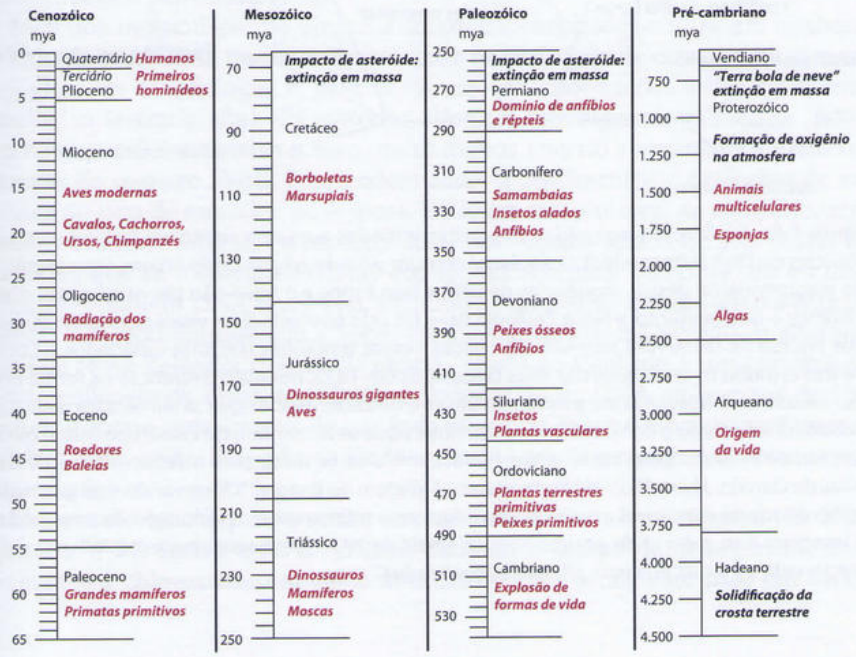
Agrupamento hierárquico é a formação de grupos de grupos de...

Filogenia é a descrição de relações biológicas, geralmente expressa como uma árvore. O estabelecimento de filogenia entre objetos *pressupõe* uma homologia e *depende* da classificação. A filogenia estabelece uma topologia de relações baseada na classificação, de acordo com a similaridade de um ou mais conjuntos de características, ou em um modelo de processo evolucionário. Em muitos casos, relações filogenéticas baseadas em características diferentes são consistentes e dão suporte umas às outras. Se características diferentes induzem relações filogenéticas inconsistentes, então todas as relações hipotéticas são duvidosas. Entretanto, note que os mesmos dados de similaridade podem ser consistentes com diferentes possibilidades de topologias ou árvores.

A escala de tempo da história da Terra

Eras geológicas (por exemplo, Cenozóica), períodos (por exemplo, Quaternário) e eventos cataclísmicos (por exemplo, impacto de asteróide: extinção em massa) estão representados em preto. A primeira aparição ou prevalência de diferentes formas de vida, em vermelho.

mya = milhões de anos atrás (do inglês: *millions of years ago*)



Os resultados de análises filogenéticas geralmente são apresentados como uma árvore evolucionária. A taxonomia dos ratitas – grandes aves não voadoras, como ema e avestruz – é um exemplo típico (Figura 4.8a). Acredita-se que o ascendente dos ratitas seja uma ave que podia voar, provavelmente relacionada com as aves tinamiformes existentes (por exemplo, a codorna).

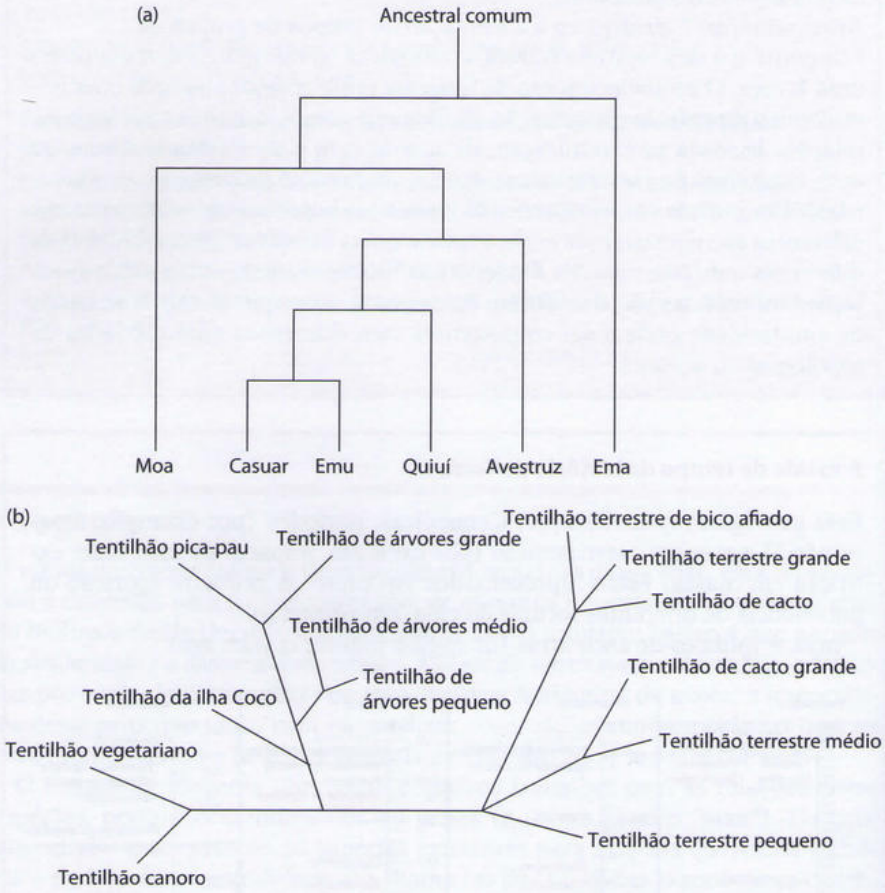


Figura 4.8 (a) Árvore filogenética dos ratitas (grandes aves não voadoras) baseada em seqüências de DNA mitocondrial. O ancestral comum situa-se na *raiz* desta árvore. Uma implicação surpreendente dessas seqüências de DNA é que a moa e o quiuí não são os parentes mais próximos e que, portanto, a Nova Zelândia deve ter sido povoada duas vezes por ratitas ou por seus ancestrais. (b) Árvore *sem raiz* das relações entre tentilhões das ilhas Galápagos e Coco. Darwin estudou os tentilhões das ilhas Galápagos em 1835, notando diferenças na forma dos seus bicos e a correlação entre a forma do bico e a dieta. Tentilhões que se alimentam de frutas possuem bicos como o dos papagaios e tentilhões que se alimentam de insetos possuem bicos mais finos e incisivos. Estas observações foram como uma semente para o desenvolvimento das idéias de Darwin. Já em 1839 ele escreveu em 'A Viagem do Beagle': "Observando esta gradação e diversidade de estruturas em um grupo pequeno e intimamente relacionado de aves, pode-se imaginar que, a partir de poucas aves originais deste arquipélago, uma espécie havia sido selecionada e modificada para diferentes finalidades".

Este tipo de árvore, mostrando todos os descendentes de uma única espécie ancestral original, é chamado de **árvore com raiz**. (A raiz da árvore geralmente aparece no topo ou ao lado; os botânicos terão que se acostumar com isso!) Alternativamente, podemos ser capazes de especificar as relações, mas não ordená-las de acordo com a história. A relação entre os tentilhões das ilhas Galápagos, estudados por Darwin, juntamente com uma espécie relacionada da ilha Coco vizinha, são mostradas em uma **árvore sem raiz** (Figura 4.8b). A adição de dados de espécies continentais da América do Sul, ancestrais dos tentilhões das ilhas, nos permitiria atribuir uma **raiz à árvore**.

O estabelecimento de uma árvore de relações pode revelar apenas a conectividade ou topologia da árvore, em cujo caso os comprimentos dos ramos não contêm informações. Um objetivo mais audacioso é mostrar a distância entre os taxa de forma quantitativa, por exemplo, rotulando os ramos com o tempo desde a divergência de um ancestral comum.

Dado um conjunto de informações que caracterizem diferentes grupos de organismos – por exemplo, seqüências de DNA ou de proteínas, ou estruturas de proteínas, ou formato dos dentes de diferentes espécies de animais – como podemos derivar informações sobre a relação entre os organismos nos quais estas características foram observadas? É raro que as relações entre as espécies e a ancestralidade sejam observáveis diretamente. Árvores evolucionárias determinadas com dados genéticos são freqüentemente baseadas em inferências a partir de padrões de similaridade, o que é tudo que pode ser observado hoje entre as espécies vivas. Geralmente assumimos que, quanto mais similares forem as características, mais relacionadas serão as espécies, apesar de esta ser uma asserção perigosa. Contudo, a partir das relações entre as características, desejamos inferir padrões de ancestralidade: a *topologia* das relações filogenéticas (de maneira informal, a "árvore genealógica").

Até que ponto as topologias das relações dependem da escolha de caráter? Em particular, existem discrepâncias *sistemáticas* entre as implicações de análises moleculares e paleontológicas?

Métodos moleculares de filogenia foram desenvolvidos contra um conhecimento de taxonomia tradicional, com base na variedade de características morfológicas, na embriologia, e, para os fósseis, em informações sobre o contexto geológico (estratigrafia). Os métodos clássicos têm algumas vantagens. Taxonomistas tradicionais têm acesso muito menos restrito a organismos extintos, através do registro fóssil. Eles podem *datar* aparecimentos e extinções de espécies através de métodos geológicos. Biólogos moleculares, ao contrário, têm acesso bastante limitado a espécies extintas. Alguns restos de subfósseis de espécies que se tornaram extintas em épocas recentes, como há um ou dois séculos, possuem DNA decifrável, incluindo espécies de quaga (um parente da zebra) e o lobo-da-austrália ("lobo" da Tasmânia, um marsupial), e algumas aves da Nova Zelândia (incluindo moas). Vimos no Capítulo 1 um exemplo de uma seqüência de mamute. Algumas seqüências de DNA de homens de Neandertal foram recuperadas de um indivíduo que morreu a aproximadamente 30.000 anos atrás. Mas não se iluda, o *Parque dos Dinossauros* continua sendo uma ficção!

Um evento crucial para o aceite dos métodos moleculares ocorreu em 1967 quando V. M. Sarich e A. C. Wilson dataram o tempo de divergência entre humanos e chimpanzés em cinco milhões de anos atrás, com base em dados

imunológicos. Naquela época, paleontólogos haviam datado esta divisão em 15 milhões de anos atrás e estavam relutantes em aceitar o método molecular. A reinterpretação do registro fóssil levou à aceitação da data de divisão mais recente e rompeu a barreira contra a aceitação de métodos moleculares. (Atualmente é aceito que humanos e chimpanzés divergiram entre aproximadamente seis e oito milhões de anos atrás.)

De fato, muitas propriedades moleculares têm sido utilizadas em estudos filogenéticos, algumas surpreendentemente já há bastante tempo. Testes sorológicos de reação cruzada foram utilizados desde o início do século passado até serem suplementados pelo uso direto de seqüências. Em um dos estudos científicos mais prematuros que eu conheço, E. T. Reichert e A. P. Brown publicaram, há quase um século (em 1909), uma análise filogenética de peixes baseada em cristais de hemoglobina. Este trabalho foi baseado na lei de Stenö (1669), que postula que apesar de cristais diferentes da mesma substância terem dimensões diferentes – alguns são grandes, outros pequenos – eles possuem os mesmos ângulos interfaciais, refletindo a similaridade no arranjo e empacotamento microscópico das unidades atômicas ou moleculares nos cristais. Reichert e Brown mostraram que os ângulos interfaciais dos cristais de hemoglobinas isolados a partir de diferentes espécies apresentavam padrões de similaridade e divergência paralelos às relações taxonômicas entre as espécies.

Os resultados de Reichert e Brown estão repletos de implicações significativas. Eles mostraram que proteínas têm formas definidas e constantes, uma idéia de forma alguma reconhecida naquela época. Eles estabeleceram que, conforme as espécies progressivamente divergem, as estruturas de suas hemoglobinas também divergem progressivamente. Em 1909 não se tinha a menor idéia sobre seqüências de ácidos nucléicos ou de proteínas. Em princípio, portanto, o reconhecimento da evolução de estruturas de proteínas precedeu, em muitas décadas, a idéia de evolução de seqüências.

Atualmente, seqüências de DNA fornecem a melhor medida de similaridades entre espécies para análises filogenéticas. Os dados são digitais. É até mesmo possível distinguir entre alterações genéticas seletivas e não-seletivas utilizando a terceira posição dos códons, ou regiões não traduzidas, tais como pseudogenes ou a proporção de substituições sinônimas e não-sinônimas de códons. Muitos genes estão disponíveis para comparação. Isso é ótimo, pois dado um conjunto de espécies a serem estudadas, é necessário encontrar genes que variem a uma taxa apropriada. Genes que se mantêm quase constantes entre as espécies de interesse não fornecem nenhuma discriminação dos graus de similaridade. Genes que variam muito não podem ser alinhados. Existe uma situação análoga na datação radioativa, que requer a escolha de um isótopo cuja meia-vida tem a mesma magnitude geral que o intervalo de tempo a ser determinado.

Por sorte, os genes variam grandemente nas suas taxas de mudanças. O genoma mitocondrial de mamíferos, uma molécula de DNA dupla fita circular, de aproximadamente 16.000 pb, provê um conjunto de seqüências que muda rapidamente e é útil para o estudo da evolução entre espécies estreitamente relacionadas. Em contraste, seqüências de RNA ribossomal foram utilizadas por C. Woese para identificar os três reinos principais: Archaea, Bacteria e Eukarya (ver Figura 1.2).

Por outro lado, diferentes taxas de mudanças nas seqüências de diferentes genes podem levar a resultados diferentes, e até mesmo contraditórios, em estudos filogenéticos. Isso é particularmente verdadeiro se desejarmos não apenas a topologia das relações, mas também os comprimentos dos ramos. Além disso, a transferência horizontal de genes e a evolução convergente são fenômenos competitivos – isto é, que competem com a descendência – que interferem com a dedução de relações filogenéticas. Na tentativa de determinar as relações filogenéticas cada vez mais cedo na escala de tempo, descobriu-se que a transferência horizontal de genes parece ter sido mais importante nos primeiros estágios da história da vida.

Árvores filogenéticas

Nós descrevemos as relações filogenéticas como árvores. Na ciência da computação, uma árvore é um tipo particular de grafo. Um grafo é uma estrutura que contém nós (pontos abstratos) conectados por arestas (representadas como linhas entre os pontos) (ver Quadro Glossário de termos relacionados a grafos; desenvolveremos a idéia de grafos em conexão com a discussão de redes no Capítulo 6). Um **caminho** entre um nó e outro é um conjunto de arestas consecutivas, começando em um ponto e terminando em outro, como a nossa viagem de Malmö a Tromsø. Um **grafo conexo** é um grafo contendo ao menos um caminho entre dois pontos quaisquer. A partir desses conceitos, podemos definir uma **árvore**: um grafo conexo, no qual há *exatamente* um caminho entre cada dois pontos. Um nó particular pode ser selecionado como a **raiz**, mas isso não é necessário – árvores abstratas podem ter ou não raiz (ver Figura 4.8). Árvores sem raiz mostram a topologia da relação, mas não o padrão de descendência. Uma árvore com raiz, na qual cada nó tem dois descendentes, é chamada de **árvore binária** (ver Programa PERL na página 224).

Outro tipo especial de grafo é o **grafo orientado**, onde cada aresta é uma via de mão única (ver página 335). Exemplos incluem os modelos ocultos de Markov (HMMs) mostrados na Figura 4.6 e as redes neurais ilustradas no Capítulo 5. Árvores filogenéticas com raiz são, implicitamente, grafos orientados com a relação ancestral-descendente determinando o sentido de cada aresta.

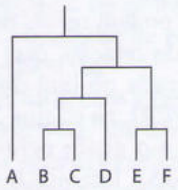
Pode ser possível atribuir números às arestas de um grafo para significar, de algum modo, uma "distância" entre os nós conectados pela aresta. O grafo pode então ser desenhado em escala, com o tamanho das arestas proporcional aos comprimentos atribuídos. O comprimento de um caminho através do grafo é a soma dos comprimentos das arestas.

Em árvores filogenéticas, os comprimentos das arestas significam uma medida da dissimilaridade entre duas espécies ou o tempo decorrido desde a sua separação. A hipótese de que diferenças entre propriedades de organismos vivos refletem seus tempos de divergência somente será verdadeira se as taxas de divergência forem as mesmas em todos os ramos da árvore. Muitas exceções são conhecidas. Por exemplo, entre os mamíferos, muitas proteínas de roedores apresentam taxas evolucionárias relativamente rápidas (ver *Weblema* 4.8).

Programa PERL 4.2 Um programa para desenhar árvores binárias

```
#!/usr/bin/perl
#drawtree.prl -- draws binary trees (root at top)
#usage:  echo '(A((BC)D)(EF))' | drawtree.prl > output.ps

print <<EOF;
%!PS-Adobe-\n%%BoundingBox: atend
/n /newpath load def /m /moveto load def /l /lineto load def
/rm /rmoveto load def /rl /rlineto load def /s /stroke load def
1.0 setlinewidth 50 100 translate 2 2 scale
/Helvetica findfont 10 scalefont setfont
EOF

$tree = <>; chop($tree); $_ = reverse($tree); s/[()]//g;

$x = 0; $y = 0;
while ($nd = chop()) {
   print "$x $y m ($nd) stringwidth pop -0.5 mul 0 rm ($nd) show\n";
   $xx{$nd} = $x; $x+=20; $yy{$nd} = 10;
}

while ($tree =~ s/\(?([A-Z])([A-Z])\)?/\1/) {
    print "n $xx{$1} $yy{$1} m\n";
    ($yy{$1} > $yy{$2}) || ($yy{$1} = $yy{$2}); $yy{$1} += 20;
    print "$xx{$1} $yy{$1} l $xx{$2} $yy{$1} l $xx{$2} $yy{$2} l s\n";

    $xx{$1} = 0.5*($xx{$1} + $xx{$2});
}
print "n $xx{$tree} $yy{$tree} m 0 20 rl s showpage\n";

$rx = 2*$x + 30; $yt = 2*$yy{$tree} + 146;
print "%%BoundingBox: 40 95 $rx $yt\n";
```

Dados de entrada: (A((BC)D)(EF)) produz o arquivo de saída abaixo, em formato PostScript, que pode ser impresso na maioria das impressoras e visualizado na maioria dos terminais.



Glossário de termos relacionados a grafos

Grafo é uma estrutura abstrata contendo nós (pontos) e arestas (linhas que conectam os pontos).
Caminho é um conjunto de arestas consecutivas.
Grafo conexo é um grafo no qual existe ao menos um caminho entre cada dois nós.
Árvore é um grafo conexo com exatamente um caminho entre cada dois nós.
Comprimento da aresta é um número atribuído a cada aresta e que significa, de alguma forma, a distância entre os nós conectados pela aresta.
Comprimento do caminho é a soma dos comprimentos das arestas que formam o caminho.

De forma geral, existem duas abordagens para a derivação de árvores filogenéticas. Uma abordagem não faz referência a qualquer modelo histórico das relações. Ela funciona medindo um conjunto de distâncias entre espécies e gera uma árvore pelo método de agrupamento hierárquico. Ela é chamada de método

fenético. A abordagem alternativa, denominada cladística, considera possíveis rotas de evolução, infere as características do ancestral em cada nó e escolhe a árvore ótima de acordo com algum modelo de mudança evolucionária. A fenética é baseada na similaridade, enquanto a cladística é baseada na genealogia.

Métodos de agrupamento

Abordagens fenéticas, ou agrupamento, para a determinação de relações filogenéticas, são explicitamente não-históricas. De fato, o agrupamento hierárquico é perfeitamente capaz de gerar uma árvore mesmo na ausência de relações evolucionárias. Uma loja de departamentos possui objetos agrupados em seções de acordo com o tipo de produto – por exemplo, vestuário ou móveis – e subgrupados em subdepartamentos mais estreitamente relacionados, tais como calçados masculinos e femininos. Calçados masculinos e femininos possuem um ancestral comum, mas não há sugestões de que calçados e móveis também o tenham.

Um procedimento de agrupamento simples funciona da seguinte maneira: dado um conjunto de espécies, determine para todos os pares uma medida da similaridade ou diferença entre elas. Isso pode depender de características físicas corporais, como a diferença entre a altura adulta média de membros de duas espécies. Ou é possível utilizar o número de bases diferentes em um alinhamento de seqüências de DNA mitocondrial. Para gerar uma árvore a partir de um conjunto de dissimilaridades, primeiro escolha as duas espécies mais parecidas e insira um nó para representar o seu ancestral comum. Depois substitua as duas espécies selecionadas por um conjunto contendo ambas as espécies e substitua as distâncias do par em relação aos outros pela média das distâncias das duas espécies selecionadas em relação às outras. Agora temos um conjunto de dissimilaridades par a par, não entre espécies individuais, mas entre conjuntos de espécies (considere cada espécie individual restante como um conjunto contendo apenas um elemento). Então, repita o processo como no exemplo a seguir.

Exemplo 4.7

Considere quatro espécies caracterizadas pelas seqüências homólogas ATCC, ATGC, TTCG e TCGG. Considerando o número de diferenças como a medida da dissimilaridade entre cada par de espécies, utilize um procedimento de agrupamento simples para derivar uma árvore filogenética.

A matriz de distâncias é:

	ATCC	ATGC	TTCG	TCGG
ATCC	0	**1**	2	4
ATGC		0	3	3
TTCG			0	2
TCGG				0

Como a matriz é simétrica, precisamos preencher apenas a metade superior. A menor distância é 1 (em negrito) entre ATCC e ATGC. Portanto, nosso primeiro grupo é {ATCC, ATGC}. A árvore conterá o fragmento:

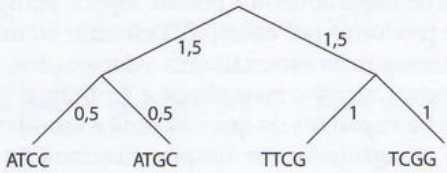
→

Exemplo 4.7 (*continuação*)

A matriz de distância reduzida é:

	{ATCC, ATGC}	TTCG	TCGG
{ATCC, ATGC}	0	$\frac{1}{2}(2 + 3) = 2{,}5$	$\frac{1}{2}(4 + 3) = 3{,}5$
TTCG		0	**2**
TCGG			0

O próximo grupo é {TTCG, TCGG}, com distância igual a **2**. Por fim, a união dos grupos {ATCC, ATGC} e {TTCG, TCGG} origina a árvore:



Os comprimentos dos ramos foram atribuídos de acordo com a regra: comprimento do ramo da aresta entre os nós X e Y = ½ da distância entre X e Y.

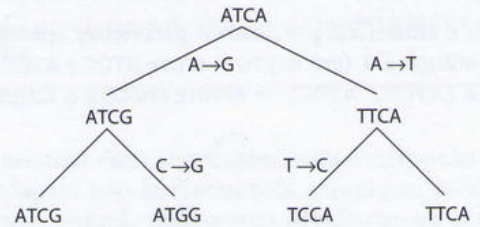
Se os comprimentos dos ramos são ou não realmente proporcionais aos tempos de divergência dos taxa representados pelos nós é uma questão a ser determinada por evidências externas.

Este processo de construção de árvores é chamado de método UPGMA (do inglês Unweighted Pair Group Method with Arithmetic Mean). Uma modificação do método UPGMA realizada por N. Saitou e M. Nei, chamada de junção de vizinhos (*Neighbour Joining*), é programada para corrigir taxas desiguais de evolução em ramos distintos da árvore.

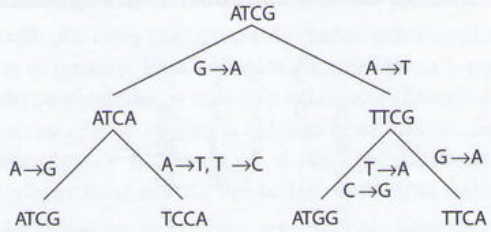
Métodos cladísticos

Métodos cladísticos lidam explicitamente com padrões de ancestralidade sugeridos pelas árvores possíveis que relacionam um conjunto de taxa. Seu objetivo é selecionar a árvore correta pela utilização explícita de um modelo de processo evolucionário. Os métodos cladísticos mais populares na filogenia molecular são a **máxima parcimônia** e a **máxima verossimilhança**. Esses métodos são especializados em dados de seqüências, começando com um alinhamento múltiplo de seqüências. Tanto a máxima parcimônia quanto a máxima verossimilhança não poderiam ser aplicadas a características anatômicas, como a altura adulta média.

O método de *máxima parcimônia* de W. Fitch define uma árvore ótima como aquela que pressupõe o menor número de mutações. Por exemplo, dada as espécies caracterizadas pelas seqüências homólogas ATCG, ATGG, TCCA e TTCA, a árvore:



pressupõe quatro mutações. A árvore alternativa:



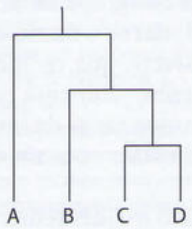
pressupõe sete mutações. Note que a segunda árvore implica que a mutação G → A na quarta posição ocorreu duas vezes, de forma independente. A primeira árvore é a ótima de acordo com o método de máxima parcimônia, pois nenhuma outra árvore envolve menos mutações. Em muitos casos, diversas árvores podem pressupor o mesmo número de mutações, sendo este número menor do que em qualquer outra árvore. Nestes casos, o método de máxima parcimônia não fornece uma resposta única.

O método de *máxima verossimilhança* atribui probabilidades quantitativas a eventos mutacionais, em vez de apenas contá-los. Assim como a máxima parcimônia, a máxima verossimilhança reconstrói ancestrais para todos os nós de cada árvore considerada; mas também atribui comprimentos aos ramos, com base nas probabilidades dos eventos de mutação pressupostos. Para cada topologia de árvore possível, as taxas de substituição assumidas são variadas para encontrar os parâmetros que forneçam a maior probabilidade de gerar as seqüências observadas. A árvore ótima é aquela que apresentar a maior probabilidade (mais verossímil) de gerar os dados observados.

Ambos os métodos, de máxima parcimônia e de máxima verossimilhança, são superiores às técnicas de agrupamento. Isso foi demonstrado em casos onde evidências independentes – por exemplo, da paleontologia – fornecem uma resposta correta, e também com dados simulados – geração computacional de seqüências em evolução.

O problema da variação das taxas de evolução

Suponha que quatro espécies, A, B, C e D, possuam a seguinte árvore filogenética:



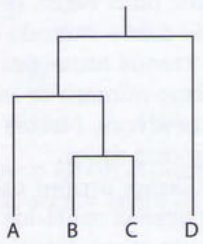
Esta árvore é consistente com a matriz de dissimilaridade:

	A	B	C	D
A	0	3	3	3
B		0	2	2
C			0	1
D				0

Suponha, no entanto, que o táxon D está mudando muito rapidamente, embora a filogenia permaneça inalterada. A matriz de dissimilaridade então pode ser:

	A	B	C	D
A	0	3	3	20
B		0	2	20
C			0	20
D				0

a partir da qual derivaríamos a árvore filogenética incorreta:



Todos os métodos discutidos aqui estão sujeitos a erros deste tipo se as taxas de mudança evolucionária variarem entre os diferentes ramos da árvore. Para testar taxas de variação, compare as espécies em consideração com um grupo externo (*outgroup*) – uma espécie mais distantemente relacionada às espécies em questão do que qualquer par destas espécies uma contra as outras. Por exemplo, se estivermos estudando espécies de primatas, um mamífero não primata como a vaca pode ser um grupo externo conveniente. Se as taxas de evolução entre as espécies de primatas fossem constantes, esperaríamos observar uma medida de dissimilaridade mais ou menos constante entre todas as espécies de primatas e a vaca. Se isso não for observado, sugere-se que as taxas de evolução variaram entre os primatas, e que a característica sendo utilizada pode não fornecer a árvore filogenética correta.

Considerações computacionais

Os métodos cladísticos – máxima parcimônia e máxima verossimilhança – são mais acurados do que os métodos de agrupamento simples, como UPGMA, mas requerem mais tempo computacional se o número de espécies for considerável. O número total de árvores possíveis, que os métodos cladísticos estão comprometidos a considerar, se puderem, aumenta rapidamente com o número de espécies. Como resultado, em muitos casos de interesse, esses métodos podem fornecer apenas respostas aproximadas, mesmo em relação às suas suposições intrínsecas.

Como as filogenias calculadas são freqüentemente aproximações, é importante tentar testá-las. Estes métodos incluem:

1. Comparações de filogenias obtidas com diferentes características que descrevem o mesmo conjunto de taxa – os resultados são consistentes? Se árvores geradas a partir de características diferentes compartilham uma mesma subárvore, talvez aquela parte da filogenia tenha sido determinada de maneira confiável, e as outras partes não.
2. Análises de subconjuntos de taxa devem fornecer a mesma resposta – em relação ao subconjunto – obtida com a análise da árvore total.

3. Testes formais de estatística, envolvendo reexecução de cálculos em subconjuntos dos dados originais, são conhecidos como *jackknifing* e *bootstrapping*:

- *Jackknifing* é o cálculo com conjuntos de dados amostrados aleatoriamente a partir dos dados originais. Para cálculos filogenéticos a partir de alinhamentos múltiplos, de seqüências, o método seleciona diferentes subconjuntos de posições no alinhamento e realiza o cálculo novamente. Se cada subconjunto fornecer a mesma árvore filogenética, a sua credibilidade aumenta. Se cada subconjunto fornecer uma árvore diferente, nenhuma delas é confiável.
- *Bootstrapping* é similar ao *jackknifing*, exceto pelo fato de que as posições escolhidas aleatoriamente podem incluir múltiplas cópias da mesma posição para formar conjuntos de dados com o mesmo tamanho do conjunto original, a fim de preservar as propriedades estatísticas da amostragem.

4. Se houver arestas muito longas, considere seriamente a possibilidade de variações desiguais na taxa evolucionária que podem ter perturbado os cálculos. Introduza um grupo externo para averiguar.

Recursos na *web*: Árvores filogenéticas

SÍTIO DA WEB ASSOCIADO

A comunidade taxonômica concentrou grandes esforços para produzir *softwares* usáveis e confiáveis. O pacote PHYLIP (do inglês PHYLogeny Inference Package) de J. Felsenstein é uma coleção integrada de muitas técnicas diferentes. Os programas funcionam em diferentes tipos de computadores e têm distribuição gratuita e são de fácil obtenção.

Sumários de ferramentas filogenéticas, incluindo listas úteis de sítios na *web*, e uma listagem geral de programas para filogenias:

`http://evolution.genetics.washington.edu/phylip/software.html`

e Whelan, S., Liò, P. & Goldman, N. (2001), Molecular phylogenetics: State-of-the-art methods for looking into the past, *Trends in Genetics*, 17, 262-272.

Alguns pacotes para o alinhamento múltiplo de seqüências, como o CLUSTAL-W, fornecem recursos para iniciar um cálculo de árvore filogenética a partir dos alinhamentos que produz.

Leituras recomendadas

Altschul, S. F. & Koonin, E. V. (1998), Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases, *Trends Biochem. Sciences*, **23**, 444-447. [Descrição de uma das ferramentas mais importantes para buscas em bancos de dados por similaridade entre seqüências.]

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994), Issues in searching molecular sequence databases, *Nature Genetics*, **6**, 119-129. [Revisão geral sobre os desafios no desenvolvimento de métodos de recuperação de informações e interpretação de resultados.]

Eddy, S. (1996), Hidden Markov Models, *Current Opinion in Struct. Biol.*, **6**, 361-365. [Leitura introdutória sobre esta importante técnica matemática que fornece ferramentas poderosas para a detecção de seqüências distantemente relacionadas e o reconhecimento de enovelamento de proteínas.]

Efron, B. & Gong, G. (1983), A leisurely look at the bootstrap, the jackknife, and cross-validation, *The American Statistician*, **37**, 36-48. [Artigo clássico sobre métodos estatísticos para a calibração de procedimentos para o reconhecimento de padrões.]

Gura, T. (2000), Bones, molecules . . . or both? *Nature*, **406**, 230-233. [Discussão sobre as congruências e os conflitos entre árvores evolucionárias baseadas em moléculas e em morfologia.]

Penny, D., Hendy, M. D., Zimmer, E. A & Hamby, R. K. (1990), Trees from sequences: Panacea or Pandora's box? *Aust. Syst. Bot.*, **3**, 21-38. [Notas sobre precauções na determinação de árvores filogenéticas.]

Whelan, S., Liò, P. & Goldman, N. (2001), Molecular phylogenetics: State-of-the-art methods for looking into the past, *Trends in Genetics*, **17**, 262-272. [Revisão com conexões para *softwares*.]

Exercícios, problemas e *weblemas*

Exercícios

4.1 Qual é a distância de Hamming entre as palavras DECLENSION e RECREATION?

4.2 Qual é a distância de Levenshtein entre as palavras BIOINFORMATICS e CONFORMATION?

4.3 A distância de Levenshtein entre a seqüência `agtcc` e `cgctca` é igual 3, consistente com o seguinte alinhamento:

```
ag-tcc
cgctca
```

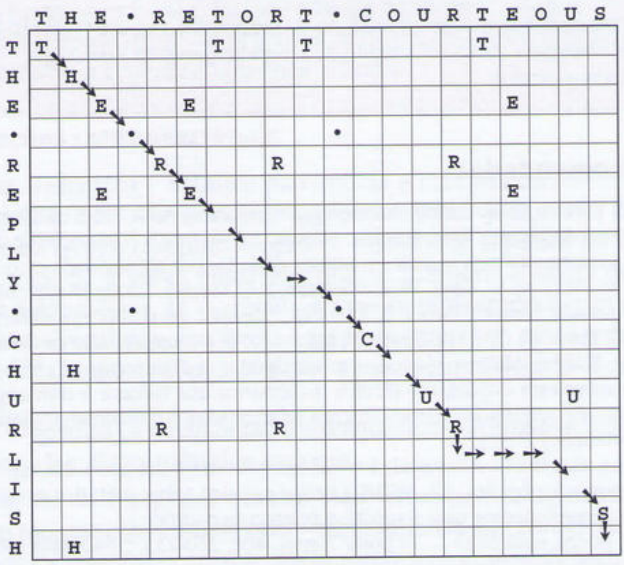
Forneça uma seqüência de três operações de edição que converta `agtcc` em `cgctca`.

4.4 *"I wasted time and now doth time waste me."* (a) Faça um esboço da aparência esperada do gráfico de pontos desta seqüência de caracteres contra ela mesma. (b) Calcule o gráfico de pontos exatamente, marcando apenas caracteres idênticos como pontos na matriz e compare com (a).

4.5 Que valores de janela e limiar (ver programa, página 184) você utilizaria para eliminar as identidades isoladas *(singletons)* no gráfico de pontos DOROTHYHODGKIN, porém mantendo as outras identidades mostradas?

4.6 Para cada uma das matrizes (a) PAM250 e (b) BLOSUM62, qual substituição é mais provável, W ↔ F ou H ↔ R?

4.7 O caminho através do gráfico de pontos mostrado corresponde a qual alinhamento?



4.8 No planejamento da sua viagem de Malmö a Tromsø (ver página 197), suponha que por motivos pessoais você desejasse incluir uma vista a Uppsala. Como você poderia ajustar os custos dos segmentos para garantir que a rota de custo mínimo passe por Uppsala?

4.9 Como você utilizaria um gráfico de pontos para reconhecer seqüências de DNA palindrômicas do tipo que aparece parcialmente em cada fita, como nos sítios de especificidade de ligação das enzimas endonucleases?

4.10 Modifique o programa PERL da página 184 que desenha gráficos de pontos para aceitar seqüências no formato FASTA.

4.11 A qual valor *P* corresponderia um escore *Z* igual a 1 em uma distribuição normal?

4.12 Para cada um dos alinhamentos da Figura 4.2, estabeleça se estão na zona de incerteza, se são mais similares do que a zona de incerteza ou menos similares do que a zona de incerteza.

4.13 A Figura 4.2a mostra o alinhamento da seqüência da papaína do mamão (papaia) e da actinidina da fruta quiuí e o gráfico de pontos correspondente. O alinhamento das seqüências mostra dois locais onde um ou mais resíduos são deletados da seqüência da papaína e um local onde um resíduo é deletado da seqüência da actinidina. Em uma fotocópia da Figura 4.2a, indique no gráfico de pontos a posição destas inserções ou deleções.

4.14 Suponha que seja argumentado que a mistura aleatória de uma seqüência não seja um modo apropriado de gerar uma população-controle para a análise estatística da significância de um alinhamento de seqüências par a par, pois as seqüências naturais possuem freqüências não-aleatórias de dipeptídeos ou tripeptídeos. Qual meio aprimorado para a geração de uma população controle você sugeriria?

4.15 Comparações das seqüências de DNA de cromossomos homólogos em diferentes pessoas mostraram que, em média, 1 em cada 700 pares de base de DNA não-codificador é diferente. Noventa e cinco por cento do genoma humano é composto por DNA não-codificador. Estime o número de polimorfismos no genoma humano para se ter uma idéia do número de potenciais marcadores de DNA.

4.16 Mostre os cálculos que originaram a entrada com valor 65 no Exemplo 4.6. Qual é a significância da observação de que duas setas daí se originam?

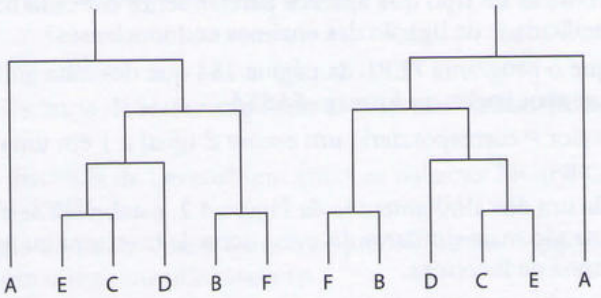
4.17 A hélice α formada pelos resíduos 32-49 na tioredoxina de *E. coli* é interrompida. Em uma fotocópia da Figura 4.5, indique onde esta interrupção ocorre. Em qual resíduo esta distorção é mais provável de ocorrer?

4.18 Em que posições da seqüência da tioredoxina de *E. coli* ocorrem voltas na estrutura que *não* estão associadas com inserções ou deleções no alinhamento de seqüências?

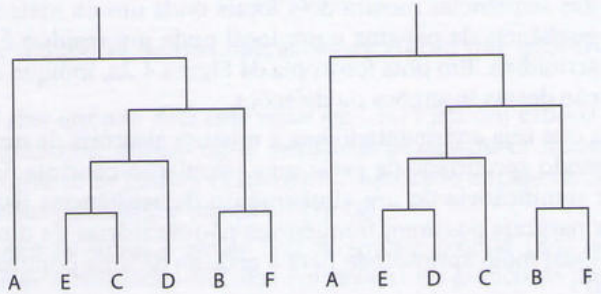
4.19 (a) Utilizando um escore de "inventário" simples, qual hexapeptídeo fornece o maior valor possível para um pareamento exato nas posições 25-30 da tabela de escores da tioredoxina (página 209)? (b) Utilizando um esquema de pontuação distribuído entre todos os 20 aminoácidos de acordo a matriz BLOSUM62, compare o escore deste hexapeptídeo com o escore do hexapeptídeo VDFSAE.

4.20 (a) Faça um inventário da região entre os resíduos 90-95, similar à tabela na página 209. Que contribuição as seqüências abaixo, alinhadas a estes resíduos, dariam a um escore de perfil simples utilizando inventários como pesos? (b) ISSAVK (c) FVGAKE.

4.21 (a) O par de árvores a seguir possui topologias idênticas?



(b) O par de árvores a seguir possui topologias idênticas?



4.22 Desenhe todas as árvores com raiz possíveis relacionando três taxa. Quantas são as árvores?

4.23 Para o grafo final no Exemplo 4.7, como foi obtido o comprimento de ramo igual a 1,5 dos nós conectando os agrupamentos {ATCC, ATGC} e {TTCG, TCGG}?

4.24 Utilizando a matriz de distâncias original e a árvore final, no Exemplo 4.7, compare, para cada par de espécies, a distância original entre elas e a soma dos comprimentos dos caminhos que as conectam na árvore.

4.25 Desenhe um exemplo de um grafo completamente conexo que não seja uma árvore.

4.26 Seqüências de DNA mitocondrial de gado europeu, africano e asiático sugerem que o rebanho europeu e o africano são mais relacionados entre si do que com o rebanho indiano. Para excluir a possibilidade do surgimento deste resultado como efeito espúrio de taxas diferenciais de evolução destas duas linhagens, sugira uma escolha razoável de uma espécie como grupo externo.

4.27 Desenhe a árvore final do Exemplo 4.7 em escala, com os tamanhos das arestas proporcionais aos comprimentos atribuídos aos ramos.

4.28 Para o método de programação dinâmica para o alinhamento de duas seqüências de comprimento n, notamos que o *tempo* de execução é proporcional a n^2. Em uma implementação simplista do algoritmo, como a necessidade de *espaço* de armazenamento dependeria de n: (a) se quisermos determinar um alinhamento ótimo, de forma que o caminho de retorno (*traceback*) precise ser armazenado? (b) se quisermos apenas o escore e não o alinhamento, de forma que o caminho de retorno não precise ser armazenado? (Nota: maneiras sutis de implementação do algoritmo diminuem substancialmente a necessidade de espaço em relação à implementação simplista.)

Problemas

4.1 Desenhe um gráfico de pontos da seqüência do genoma do vírus anão do trigo: `ttttcgtgagtgcgcggaggctttt` contra ela mesma. De que forma ela não é um palíndromo perfeito?

4.2 (a) Como você alteraria o algoritmo na seção de programação dinâmica (começando na página 196) para localizar os pareamentos ótimos de um padrão relativamente curto $A = a_1a_2...a_n$ em uma seqüência longa $B = b_1b_2...b_m$, com n ≪ m. (Nenhuma penalidade para lacunas na região B que precede e sucede a região pareada com A é permitida.) Isso corresponde à identificação de motivos descrita no Capítulo 1. (b) Refaça os cálculos do Exemplo 4.6 para o alinhamento das seqüências `ggaatgg` e B = `atg` como um problema de identificação de motivo, utilizando o mesmo esquema de escores: pareamento idêntico = 0, malpareamento = 20, introdução de lacunas *internas* = 25, extensão de lacunas = 22. (c) Como o resultado que você obteve difere do resultado derivado no exemplo?

4.3 Como você poderia modificar o método de perfis para manter sua habilidade de identificar tioredoxinas de animais não mamíferos se um grande número de mamíferos adicionais, estreitamente relacionados, fosse acrescentado à tabela? Considere (a) métodos que tentem remover a redundância ignorando determinadas seqüências e (b) métodos que retenham todas as seqüências, mas incluam um esquema de atribuição de pesos para balancear a representação das seqüências estreitamente relacionadas.

4.4 Escreva um programa PERL que faça inventários de perfis a partir de alinhamentos múltiplos de seqüências e que atribua escores para o pareamento da seqüência-alvo utilizando a matriz BLOSUM62. Assuma que a seqüência-alvo já tenha sido alinhada antes de ser fornecida ao programa.

4.5 (a) Escreva um programa PERL que leia duas seqüências de caracteres e liste todos os pareamentos de regiões de cinco caracteres contíguos. Teste o programa nestas seqüências:

```
My.care.is.loss.of.care,.by.old.care.done, e
Your.care.is.gain.of.care,.by.new.care.won
```

(b) Desenvolva este programa para estender e combinar os pareamentos encontrados com regiões mais longas que contenham pareamentos perfeitos de 5 caracteres, sem lacunas, com não mais do que 25% de desigualdade total.

4.6 Estenda o problema anterior escrevendo um programa PERL que ilustre, em uma forma baseada no gráfico de pontos, o progresso de um algoritmo do tipo BLAST nos estágios quando ele (a) detecta todas as subseqüências pareadas de comprimento igual a 5, (b) estenda-as ao máximo de pareamentos contíguos, (c) combine-as para formar um pareamento com não mais do que k desigualdades. Você pode utilizar o programa para o desenho de gráficos de pontos no texto.

4.7 Escreva um programa para *animar* o progresso de um algoritmo do tipo BLAST como descrito no problema anterior. Como fonte, procure por exemplos para animação de algoritmos de busca de seqüências na *web*. (Este problema exige um nível relativamente alto de experiência em computação.)

4.8 Cadeias de RNA fita simples, tais como tRNA, adotam conformações que contêm regiões de *haste-alça* (*stem loop*), na qual uma região da cadeia se dobra sobre ela mesma para formar uma hélice dupla fita pela complementaridade dos

pares de bases com as cadeias em disposição antiparalela. Como um programa que detectaria palíndromos seria útil na análise de seqüências de RNA para detectar regiões capazes de formar estrutura em haste-alça perfeitas (ou seja, sem bases malpareadas)?

4.9 Compare a predição dos limites de resíduos para hélices transmembrana de bacteriorrodopsinas obtida com o método Phobius com os valores observados em estruturas experimentais (Figura 4.7).

4.10 Suponha que você tenha um par de dados, um vermelho e um verde. Defina um *estado* do par de dados como o seguinte par de números: o número que aparece na face superior do dado vermelho seguido do número que aparece na face superior do dado verde. Em vez de rolar os dados, passe de um estado a outro pela rotação de cada dado em 90° em qualquer direção, com igual probabilidade. Dessa forma, um estado em que o número 6 está na face superior de um dado pode ser seguido, com igual probabilidade, por 2, 3, 4 ou 5 (dados são construídos de forma que a soma dos números em cada um dos três conjuntos de faces opostas seja 7. Assim, a probabilidade de que 1 siga 6 é zero, pois isso requer uma rotação de 180°). A probabilidade da seqüência 6, 2, 6, 4 é $(1/4)^4 = 1/256$. A probabilidade de gerar a seqüência 6, 2, 5, 4 é zero, pois a transição 2→5 não é permitida, e a probabilidade da seqüência 6, 6, 3, 4 também é zero, pois o sistema precisa mudar seu estado, então 6 não pode suceder 6.

Este procedimento define um processo de Markov de primeira ordem.

Escreva um programa que responda às seguintes questões: Suponha que o estado inicial tenha o número 4 no topo do dado vermelho e o número 3 no topo do dado verde. (a) Qual a probabilidade de outro estado, no qual os números somem 7, apareça em 5 movimentos? (b) Se o estado inicial for um 8, qual a probabilidade de que outro 8 apareça antes de um 7?

4.11 Mostre que qualquer grafo (não orientado) que tenha uma das seguintes propriedades também deve possuir a outra: (1) Existe um único caminho entre quaisquer dois nós. (2) O grafo não contém ciclos.

4.12 Quantos grafos existem ao todo entre o ponto Início e Término na Figura 4.9? Conte-os em cada uma das seguintes maneiras:

(a) Força-bruta – escreva todas as possibilidades. Isso é na verdade um exercício que exige menos intelecto do que parece. Isso demonstra que na realidade não é tão complicado quanto parece. Em segundo lugar, isso mostra como você percebe padrões enquanto o faz.

(b) Na Figura 4.9, conte o número de caminhos entre Início até A e de A até Término. Multiplique estes números para obter o número total de caminhos entre Início e Término, passando por A. Conte então o número de caminhos entre Início até B e de B até Término. Qual é a relação entre estes dois números? Multiplique estes números para obter o número total de caminhos entre Início e Término, passando por B. Compute o número total de caminhos entre Início e Término como a soma do número de caminhos entre Início e Término passando por A, B, C e D.

(c) Reconheça que ir do ponto inicial até o final requer 6 passos, incluindo exatamente três curvas à esquerda e três curvas à direita (de outra forma, você não terminará no ponto correto). Diferentes escolhas na ordem de curvas à direita e à esquerda correspondem a caminhos diferentes. Para contar o número de caminhos, reconheça que você precisa apenas decidir quantas maneiras existem para escolher os três passos, onde você quer dobrar à esquerda (pois então

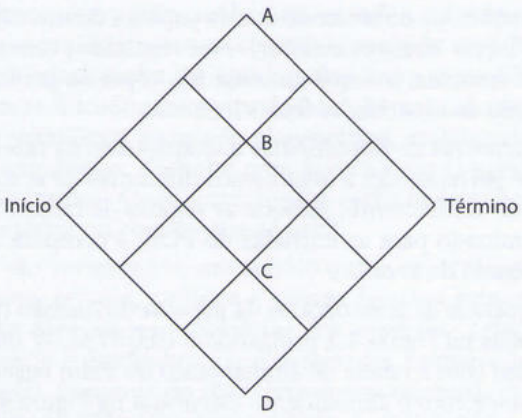


Figura 4.9 Calculando caminhos em uma rede finita.

você deve dobrar à direita nos outros três passos). Para atribuir três curvas à esquerda a seis passos, primeiro podemos escolher um dos 6 passos como uma curva à esquerda, e então um dos quatro passos restantes como a segunda curva à esquerda. No entanto, o produto destes números ultrapassa o número de possibilidades, pois inclui os mesmos conjuntos de passos atribuídos em ordens diferentes. Cada trinca pode surgir de seis formas distintas e é necessário corrigir isso. O resultado é igual ao coeficiente binomial

$$\binom{6}{3} = 6!/(3!3!).$$

4.13 Para a árvore final no Exemplo 4.7, derive possíveis ancestrais nos nós internos escolhidos pelo critério de máxima parcimônia. Existem ambigüidades?

4.14 Uma anotação conveniente para árvores utiliza parênteses combinados para indicar os agrupamentos. (a) Expanda o seguinte agrupamento em uma árvore com raiz: (A(BC)D). (b) Escreva a anotação com parênteses das árvores mostradas no Exercício 4.21.

4.15 Adicione uma quantidade adequada de comentários ao programa PERL para o desenho de árvores.

4.16 Escreva um programa PERL para o método UPGMA de derivação de árvores filogenéticas a partir de uma matriz de distâncias. Você pode utilizar o programa para o desenho de árvores para produzir uma saída gráfica.

Weblemas

4.1 Busque a seqüência do gene da subunidade 6 da ATPase mitocondrial do peixe-bruxa do Atlântico (*Myxine glutinosa*). Desenhe o gráfico de pontos contra o gene homólogo da lampreia do mar (*Petromyzon marinus*). Comente sobre a similaridade observada e compare com a similaridade observada entre as seqüências da lampreia e do tubarão pata-roxa, mostradas na página 182.

4.2 Submeta a seqüência de aminoácidos da papaína do mamão (papaia) a uma busca BLAST e PSI-BLAST. Quais homólogas mostradas na Figura 4.2 são identificadas com sucesso pelo BLAST? E pelo PSI-BLAST?

4.3 Submeta a seqüência de aminoácidos da papaína do mamão (papaia) a uma busca PSI-BLAST (ver *Weblema* anterior). Nos resultados com a identidade com a procatepsina L humana, indique em uma fotocópia do gráfico de pontos (Figura 4.2b) a região de identidades locais listadas.

4.4 Encontre estruturas de tioredoxinas que aparecem na tabela de alinhamento da Gravura IV pertencentes a organismos diferentes de *E. coli*. Em uma fotocópia da tabela de alinhamento, indique as regiões de hélices e fitas de folhas, conforme determinado para as entradas do PDB, e compare com as hélices e folhas da tioredoxina de *E. coli*.

4.5 Alinhe a seqüência de aminoácidos da papaína do mamão (papaia) e das homólogas mostradas na Figura 4.2 utilizando o CLUSTAL-W ou T-Coffee. Compare os resultados com a tabela de alinhamento no Pfam baseada nos modelos ocultos de Markov e com o alinhamento estrutural na Figura 4.2.

4.6 O PSI-BLAST pode identificar a homologia entre domínios de imunoglobulinas e os domínios da endonuclease C de *Cellulomonas fimi* e o receptor IgA de *Streptococcus agalactiae*?

4.7 (a) O PSI-BLAST pode identificar a relação entre a urease de *Klebsiella aerogenes* e a fosfotriesterase de *Pseudomonas diminuta* e a adenosina deaminase de camundongo? (b) Compare os resultados do alinhamento destas três seqüências produzidos pelos programas DALI, CLUSTAL-W e T-Coffee.

4.8 Os hormônios de crescimento na maioria dos mamíferos possuem seqüências de aminoácidos bastante similares. (O hormônio de crescimento de alpaca, cachorro, gato, cavalo, coelho e elefante diferem do hormônio dos porcos em não mais de três posições dentre 191.) O hormônio de crescimento humano é muito diferente, variando em 62 posições. A evolução do hormônio do crescimento se acelerou acentuadamente na linhagem que originou os humanos. Recuperando e alinhando seqüências de hormônio do crescimento de espécies estreitamente relacionadas aos humanos e nossos ancestrais, determine *onde*, na árvore evolutiva que originou os humanos, a aceleração da evolução do hormônio do crescimento começou.

A próxima série de *weblemas* se destina a colocar a espécie humana no seu contexto biológico pela análise de seqüências de parentes próximos e distantes e para ilustrar algumas das mais variadas informações genéticas que têm sido utilizadas para investigar relações filogenéticas.

4.9 As espécies vivas mais relacionadas aos humanos são os macacos e os chimpanzés. Os elementos Alu são um tipo de SINE (Short INterspersed Element), úteis como marcadores de espécies. Apesar de serem parte da região repetitiva não-codificadora do genoma, alguns elementos Alu atuam na regulação gênica. Com base nos elementos Alu que regulam os genes do hormônio da paratireóide, do receptor hematopoiético célula específico FcεRI-γ, do receptor α3 do sistema nervoso central específico acetilcolina nicotínica, e do CD8α específico de células T, derive uma árvore filogenética para humanos, chimpanzés, gorilas, orangotangos, babuínos, macacos-rhesus e macacos.

4.10 Humanos são primatas, uma ordem que nós, os chimpanzés e os macacos compartilhamos com os lêmures e társios. Com base no grupo de genes da β-globina de humanos, chimpanzés e macacos do velho mundo, macacos do novo mundo, lêmures e dos társios, derive uma árvore filogenética para estes grupos.

4.11 Primatas são mamíferos, uma classe que compartilhamos com os marsupiais e monotremados. Os únicos marsupiais vivem principalmente na Austrália

e nas ilhas vizinhas, exceto pelo gambá, encontrado na América do Norte e do Sul. Os monotremados existentes estão limitados a três espécies: o ornitorrinco e duas espécies relacionadas de equidna. Recupere as seqüências dos genes nucleares da manose 6-fosfato/receptor de fator do crescimento II tipo insulina para espécies de mamíferos incluindo placentários, marsupiais e monotremados. Com estas seqüências, desenhe uma árvore evolucionária, indicando os comprimentos dos ramos. Os monotremados estão mais relacionados com os mamíferos placentários ou com os marsupiais?

4.12 Mamíferos são vertebrados, um subfilo que compartilhamos com os peixes, tubarões, aves, répteis, anfíbios e peixes ágnatos primitivos (exemplo: lampreia). Derive árvores evolucionárias das espécies: celacanto (*Latimeria chalumnae*), o grande tubarão-branco (*Carcharodon carcharias*), atum saltador (*Katsuwonus pelamis*), lampreia do mar (*Petromyzon marinus*), rã-leoparda (*Rana pipiens*) e crocodilo do Nilo (*Crocodylus niloticus*), utilizando as seqüências do citocromo *c* e da ribonuclease pancreática.

4.13 Tetrápodes são gnatostômios, uma superclasse que compartilhamos com os peixes. A visão tradicional da evolução peixes → tetrápodes é que os vertebrados com mandíbula se dividiram em um grupo contendo peixes cartilaginosos, *Chondrichthyes*, incluindo tubarões e raias, e outro contendo tanto os peixes com barbatana tipo raio, *Actinopterygii*, incluindo os modernos peixes ósseos, como bacalhau e salmão, quanto os peixes com barbatana tipo lóbulo, *Sarcopterygii*, incluindo celacantos, peixes pulmonados e tetrápodes (ver Figura 4.10a). Teste esta hipótese utilizando ao menos 12 genes mitocondriais codificadores de proteínas de ao menos 30 espécies, incluindo tubarões, peixes pulmonados, peixes ósseos, anfíbios, répteis, aves e mamíferos, utilizando a lampreia como grupo externo. Dos três grupos – peixes cartilaginosos, peixes ósseos e tetrápodes – qual par parece estar mais relacionado: peixes ósseos e tetrápodes, de acordo com a visão tradicional, ou um par diferente? Considere especificamente a filogenia de acordo com a qual os tetrápodes se dividem primeiro, e os peixes cartilaginosos, peixes ósseos e mesmo os peixes pulmonados são taxa irmãos (ver Figura 4.10b).

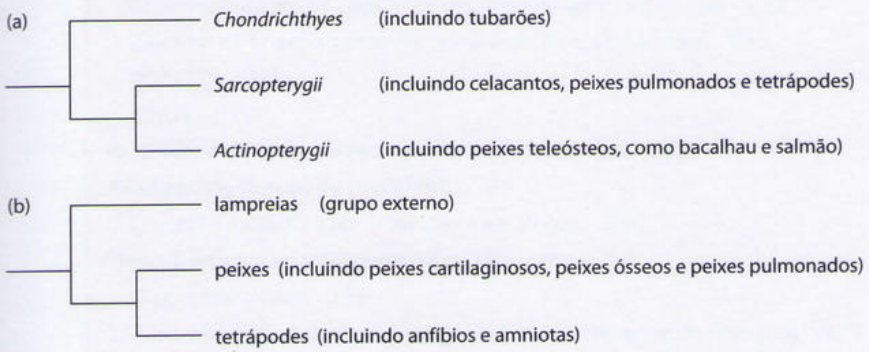


Figura 4.10 (a) Visão tradicional da evolução dos gnatostômios: a primeira divisão ocorreu entre os peixes cartilaginosos como os tubarões e outros; então os tetrápodes emergiram de um grupo que contém celacantos, peixes pulmonados e o subgrupo dos tetrápodes. Uma alternativa a ser considerada é que a divisão que origina os tetrápodes tenha ocorrido mais cedo. (b) Árvore alternativa, com raiz; utilizando a lampreia como grupo externo, onde a linhagem que origina os tetrápodes se dividiu primeiro.

4.14 Vertebrados são cordados, um filo que também inclui *lancelets* (pequenos animais marinhos que se parecem com peixes; exemplo: anfioxo) e vertebrados ágnatos (que não possuem uma coluna vertebral verdadeira; exemplo: lampreia). Como em outros organismos com simetria bilateral (incluindo os insetos), os genes HOX de vertebrados codificam uma família de proteínas de ligação ao DNA. A expressão desses genes varia ao longo do eixo corporal que percorre da cabeça à cauda e controla o estabelecimento do plano corporal. Existe, de fato, um impressionante mapeamento da ordem desses genes nos cromossomos, da ordem da sua ação ao longo do corpo e os tempos relativos do início da sua atividade durante o desenvolvimento.

Durante o curso da evolução dos vertebrados, ocorreram eventos de duplicação genômica em larga escala, associadas, presumivelmente, com o desenvolvimento de maiores complexidades de arquitetura corporal, como proposto originalmente por S. Ohno em 1970. Os genomas de insetos e do anfioxo possuem apenas um conjunto de genes HOX. O peixe "paulistinha" (*zebrafish*) possui sete conjuntos de genes HOX, interpretados em termos de uma série de duplicações: $1 \rightarrow 2 \rightarrow 4 \rightarrow 8$, seguida da perda de um, para haver a redução de $8 \rightarrow 7$.

Encontre o número de grupos de genes HOX em humanos e lampreias, faça um alinhamento múltiplo de seqüências para atribuir as correspondências entre os genes individuais e derive destes resultados uma árvore filogenética para o anfioxo, lampreia, peixes e mamíferos.

4.15 Cordados são deuterostômios (ver Figura 1.3), um agrupamento que compartilhamos com os urocordados (exemplo, seringa-do-mar), hemicordados (exemplo, anfioxo) e equidermos (exemplo, estrela-do-mar). Existem diferenças sistemáticas entre estes quatro filos no seu código genético mitocondrial. Determine, para exemplos de organismos de cada filo, os aminoácidos que correspondem aos códons ATA e AGA. Derive destes resultados uma árvore filogenética para os quatro filos deuterostômios.

CAPÍTULO 5

Estrutura de Proteínas e Descoberta de Fármacos

Sumário do capítulo

Introdução 240

Estabilidade e enovelamento de proteínas 243

O mapa de Sasisekharan-Ramakrishnan-Ramachandran descreve as conformações permitidas da cadeia principal 243

As cadeias laterais 245

Estabilidade e desnaturação de proteínas 245

Enovelamento de proteínas 248

Aplicações da hidrofobicidade 249

Sobreposição de estruturas e alinhamentos estruturais 253

DALI (Distance-matrix ALIgnment) 255

Evolução de estruturas de proteínas 256

Classificação de estruturas de proteínas 258

SCOP 259

Predição e modelagem de estrutura de proteínas 260

Critical Assessment of Structure Prediction (CASP) 262

Predição de estrutura secundária 264

Modelagem por homologia 270

Reconhecimento de padrão de enovelamento ou topologia 272

Cálculos de energia conformacional e dinâmica molecular 275

ROSETTA 279

LINUS 279

Atribuição de estruturas de proteínas a genomas 282

Predição de função de proteínas 285

Divergência de função: ortólogas e parálogas 286

Descoberta e desenvolvimento de fármacos 289

O composto líder 291

A bioinformática na descoberta e no desenvolvimento de fármacos 293

Leituras recomendadas 304

Exercícios, problemas e weblemas 305

Objetivos do aprendizado

1. Entender o conceito de enovelamento de proteínas: o processo pelo qual a seqüência unidimensional de aminoácidos, codificada por um gene, assume uma estrutura tridimensional definida e biologicamente ativa.
2. Reconhecer que considerações estéricas limitam seriamente as conformações da cadeia polipeptídica, com o mapa de Sasisekharan-Ramakrishnan-Ramachandran identificando as conformações permitidas da cadeia principal.
3. Conhecer as 20 cadeias laterais – as personagens que desempenham todas as funções em todas as proteínas.
4. Entender o efeito hidrofóbico e suas implicações na estrutura e energética de proteínas enoveladas.
5. Generalizar as idéias de alinhamento de seqüências para o alinhamento de seqüências de proteínas pela sobreposição de suas estruturas.
6. Conhecer a relação entre divergência de seqüência e divergência de estrutura na evolução de proteínas.
7. Familiarizar-se com a classificação dos padrões de enovelamento de proteínas, como, por exemplo, o apresentado pelo banco de dados Structural Classification of Proteins (SCOP) e o seu sítio na *web*.
8. Conhecer algumas abordagens básicas para a predição de estrutura de proteínas a partir da seqüência de aminoácidos, e o estado-da-arte como retratado nos projetos CASP (*Critical Assessment of Structure Prediction*).
9. Entender os fundamentos dos modelos ocultos de Markov (Hidden Markov Models), os melhores métodos disponíveis atualmente para a predição da afinidade entre proteínas, a partir de suas seqüências.
10. Conhecer os requisitos básicos de um fármaco de sucesso, e entender algumas das abordagens para o planejamento e a descoberta de fármacos.

Introdução

A grande variedade de estruturas tridimensionais e funções de proteínas surge em moléculas que compartilham características básicas comuns. Quimicamente, as proteínas são como aqueles cordões de pequenas lâmpadas coloridas em árvores de Natal: cada proteína consiste em uma cadeia principal polimérica linear (isto é, não-ramificada), com diferentes cadeias laterais de aminoácidos ligadas em intervalos regulares (Figura 1.6). O fio que liga as lâmpadas no cordão corresponde à cadeia principal, também denominada esqueleto polipeptídico, que se repete, e a seqüência variável de cores das luzes corresponde à seqüência de cadeias laterais individuais.

A seqüência de aminoácidos de uma proteína é especificada pela seqüência de nucleotídeos de um gene. As estruturas tridimensionais das moléculas de

proteínas são determinadas, sem qualquer contribuição adicional dos ácidos nucléicos, pela seqüência de seus aminoácidos. As proteínas se enovelam espontaneamente em suas conformações nativas.

Como a seqüência de aminoácidos determina a estrutura tridimensional? Qualquer enovelamento possível da cadeia principal coloca diferentes resíduos em contato. As interações das cadeias laterais e da cadeia principal, consigo mesmas e com o solvente, e as restrições existentes de mobilidade das cadeias laterais, determinam as estabilidades relativas das diferentes conformações. Isso é uma conseqüência da segunda lei da termodinâmica, a qual postula que sistemas a temperatura e pressão constantes encontram um estado de equilíbrio que é um compromisso entre o conforto (baixa entalpia, H) e a liberdade (alta entropia, S), resultando em um mínimo de energia livre de Gibbs $G = H - TS$, onde T é a temperatura absoluta. (Nos relacionamentos humanos, o casamento representa um desses compromissos.)

As proteínas evoluíram de forma que um determinado padrão de enovelamento de sua cadeia principal seja, termodinamicamente, significativamente melhor do que o de outras conformações. Este é o estado nativo. Se pudéssemos calcular, com uma precisão aceitável, as energias e entropias de diferentes conformações, e se pudéssemos analisar computacionalmente um conjunto grande dessas possíveis conformações, para termos certeza de estarmos incluindo a conformação correta, seria possível, conseqüentemente, predizer as estruturas de proteínas a partir apenas de suas seqüências de aminoácidos com base em princípios físico-químicos. Ocorreram avanços em direção à solução deste problema, mas ele ainda não foi resolvido.

A cadeia principal de cada proteína em seu estado nativo descreve uma curva no espaço. Atualmente, conhecemos as estruturas de mais de 43.000 proteínas (incluindo aquelas repetidas ou de mutantes pontuais), e identificamos nessas proteínas uma grande variedade de padrões espaciais. O primeiro problema na análise destas estruturas é o da representação. A Figura 5.1 ilustra, para a pequena proteína acilfosfatase, a dificuldade de uma interpretação completamente detalhada e de uma representação precisa, e as diversas simplificações que programas de computador utilizam para nos permitir um acesso visual a esses dados. Um ilustrador molecular habilidoso combinará esses diversos tipos de representações simplificadas para visualizar em detalhes as diferentes partes da estrutura.

A imagem central da Figura 5.1 mostra o formato espacial da cadeia principal da acilfosfatase. As duas regiões no primeiro plano das figuras têm a forma de hélices – como os pequenos postes listrados em vermelho e branco que identificam as barbearias norte-americanas – com seus eixos quase verticais, na orientação mostrada. A acilfosfatase também contém quatro fitas de uma folha β. Elas também são aproximadamente verticais nesta orientação. As quatro fitas β interagem lateralmente, para estabilizar o seu arranjo na forma de uma folha β. Na imagem inferior, as hélices e fitas são representadas como "ícones": hélices como cilindros e fitas como setas largas. A imagem superior da Figura 5.1, mostrando a representação mais detalhada da estrutura, incluindo a cadeia principal e as cadeias laterais, indica a importância da simplificação na geração de uma figura compreensível até mesmo para uma proteína pequena.

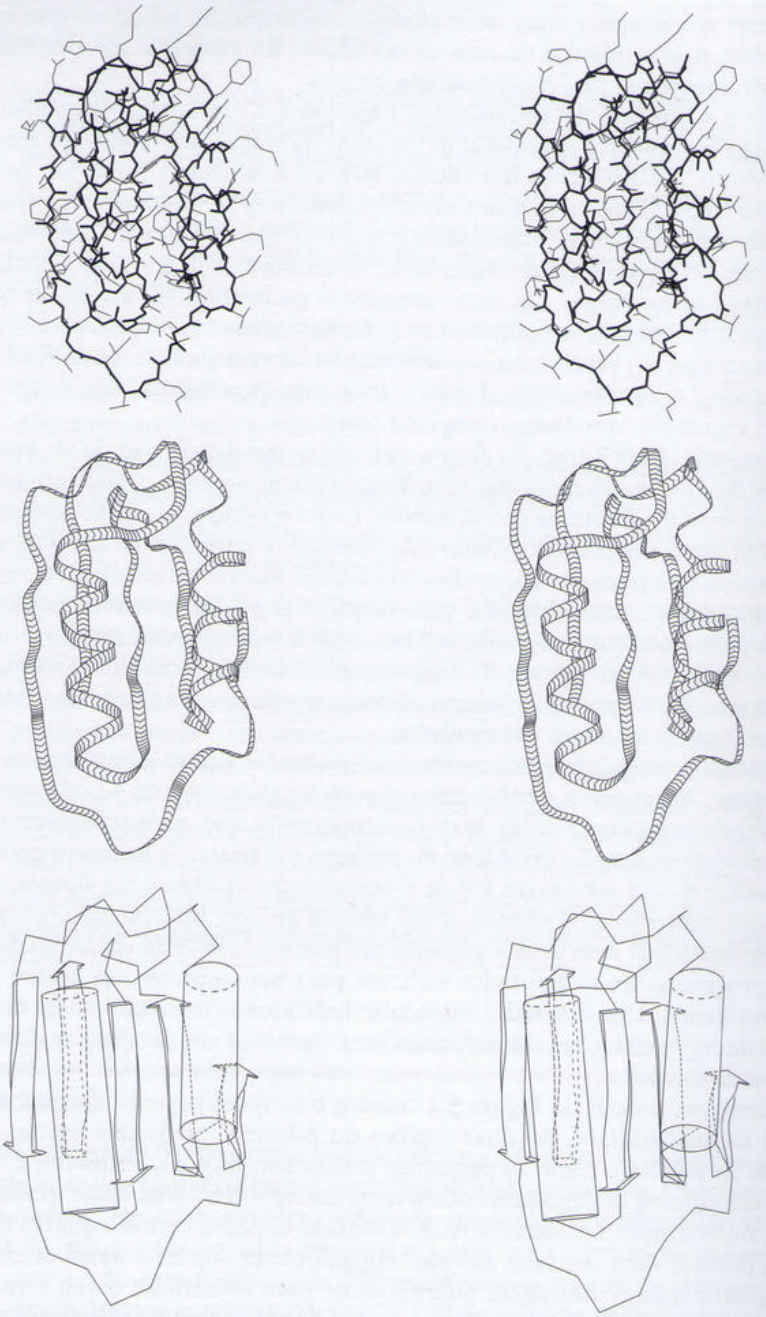


Figura 5.1 As proteínas possuem estruturas tão complexas que foi necessário o desenvolvimento de ferramentas especiais para representá-las. Esta figura mostra uma proteína relativamente pequena, a acilfosfatase, em três níveis diferentes de simplificação. Acima: modelo completo do esqueleto polipeptídico, com a cadeia principal mais destacada do que as cadeias laterais. No centro: o percurso da cadeia é representado como curvas interpoladas suaves que lembram uma fita, cujas listras anguladas indicam o sentido da cadeia. Abaixo: diagrama esquemático no qual cilindros representam as hélices e setas representam as fitas da folha β. Os objetos sólidos do desenho estão representados como "translúcidos" pela alteração das linhas que passam atrás deles para linhas tracejadas. É possível sobrepor visualmente as diferentes representações pela rotação da página em 90° e visualização em estéreo (mas não por muito tempo!).

Estabilidade e enovelamento de proteínas

Apesar de ainda não ser possível a predição da estrutura de proteínas somente a partir dos princípios básicos da física, nós entendemos a natureza geral das interações que determinam as estruturas de proteínas.

Para assumir sua estrutura nativa, a proteína deve otimizar as interações intra e inter-resíduos, sujeitas a limitações na curva espacial traçada pela cadeia principal. As conformações preferidas pela cadeia principal predispõem as formas de enovelamentos em direção a motivos estruturais recorrentes: hélices, regiões mais estendidas que interagem para compor folhas β e vários tipos de voltas.

O mapa de Sasisekharan-Ramakrishnan-Ramachandran descreve as conformações permitidas da cadeia principal

Com boa aproximação, pode-se dizer que a conformação da cadeia principal de cada resíduo de aminoácido, excluindo-se a glicina, está restrita a dois estados conformacionais discretos.

Um fragmento da cadeia polipeptídica linear, comum a todas as estruturas de proteínas, é mostrado na Figura 5.2. Rotações são permitidas em torno das ligações simples N-Cα e Cα-C de todos os resíduos (com uma exceção: prolina). Os ângulos ϕ e ψ em torno dessas ligações, e o ângulo de rotação em torno da ligação peptídica, ω, definem a conformação de um resíduo. A ligação peptídica tende a ser planar, com dois estados possíveis: *trans*, $\omega \approx 180°$ (mais comum), e *cis*, $\omega \approx 0°$ (raramente, e na maioria dos casos em um resíduo prolina). A seqüência dos ângulos ϕ, ψ e ω de todos os resíduos de uma proteína define a conformação do esqueleto polipeptídico.

O princípio de que dois átomos não podem ocupar o mesmo lugar no espaço limita os valores dos ângulos conformacionais. As faixas permitidas para ϕ e ψ, para $\omega = 180°$, situam-se nas regiões definidas de um gráfico chamado de mapa

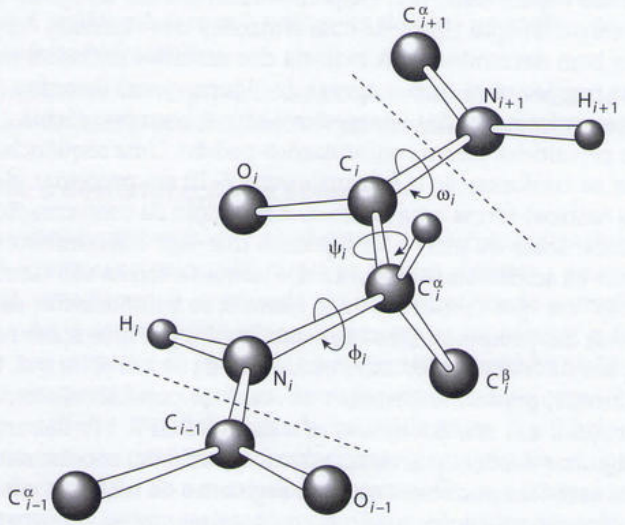


Figura 5.2 Definição dos ângulos conformacionais do esqueleto polipeptídico.

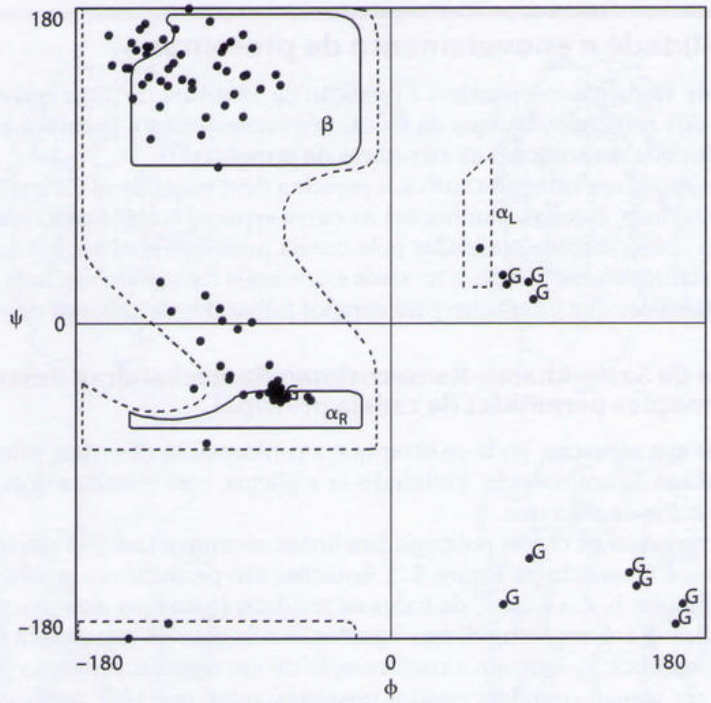


Figura 5.3 O mapa de Sasisekharan-Ramakrishnan-Ramachandran da acilfosfatase [código PDB: 2ACY]. Note a concentração de resíduos nas regiões α e β. A maioria das exceções ocorre com os resíduos de glicina (identificados por G).

de Sasisekharan-Ramakrishnan-Ramachandran – normalmente abreviado como mapa de Ramachandran (ver Figura 5.3). As linhas sólidas na figura delimitam as regiões de ϕ e ψ energeticamente preferidas; as linhas tracejadas delimitam as regiões estericamente proibidas. As conformações da maioria dos aminoácidos ficam dentro das regiões α_R ou β. A glicina tem acesso a conformações adicionais. Mais especificamente, ela pode formar hélices levógiras: α_L. A Figura 5.3 mostra a distribuição típica de conformações dos resíduos em uma estrutura protéica bem determinada. A maioria dos resíduos encontra-se dentro, ou próxima, das regiões permitidas, apesar de alguns serem forçados, pelo enovelamento, a assumirem estados energeticamente menos favoráveis.

As regiões permitidas geram conformações-padrão. Uma seqüência de resíduos consecutivos na conformação α (normalmente 6-20 em proteínas globulares em seus estados nativos) forma uma hélice α. A repetição da conformação β gera uma fita β estendida. Duas ou mais fitas β podem interagir lateralmente para formar folhas β, como na acilfosfatase (Figura 5.1). Hélices e folhas são elementos estruturais "padrão" ou "pré-fabricados" que formam os componentes das conformações da maioria das proteínas. Eles são estabilizados por interações relativamente fracas, **ligações de hidrogênio**, entre os átomos da cadeia principal. Em algumas proteínas fibrosas, praticamente todos os resíduos pertencem a um desses dois tipos de estrutura: a lã contém hélices α; a seda, folhas β. Fibrilas amilóides, formadas em algumas doenças por várias proteínas, também contêm extensas folhas β. Interações estéricas permitem que um segmento de resíduos consecutivos se localize totalmente na conformação α_R ou totalmente na conformação β, mas proíbem uma conformação de hélice seguida de uma fita e vice-versa.

Proteínas globulares típicas contêm várias regiões de hélices e/ou folhas, conectadas por *voltas*. Normalmente, as regiões terminais de uma hélice ou fita localizam-se na superfície de um domínio de uma estrutura protéica. Elas são conectadas por voltas ou alças – regiões nas quais as cadeias mudam de direção, dirigindo-se para dentro da estrutura. A maioria das voltas, mas não todas, são curtas e localizadas em uma região exposta à superfície que tende a conter resíduos carregados ou polares.

Como a cadeia principal escolhe entre as possíveis conformações permitidas? O que é único em cada proteína é a sua seqüência de cadeias laterais. Logo, as interações que envolvem as cadeias laterais devem determinar a conformação da cadeia principal.

As cadeias laterais

As cadeias laterais oferecem a versatilidade físico-química necessária para gerar todos os padrões de enovelamento diferentes. As cadeias laterais dos vinte aminoácidos variam em:

- Tamanho A menor, a glicina, consiste em apenas um átomo de hidrogênio; uma das maiores, a fenilalanina, contém um anel benzênico.
- Carga elétrica Algumas cadeias laterais possuem uma carga efetiva positiva ou negativa em pH neutro. Asp e Glu são negativamente carregadas. Lis e Arg são positivamente carregadas. (Resíduos carregados com cargas opostas podem formar um par iônico com interações atrativas chamadas de pontes salinas.)
- Polaridade Algumas cadeias laterais são polares; elas podem formar ligações de hidrogênio com outras cadeias laterais polares, ou com a cadeia principal, ou com a água. Outras cadeias laterais são eletricamente neutras. Algumas delas contêm grupos químicos relacionados com hidrocarbonetos comuns, como metano ou benzeno. Por causa da interação termodinamicamente desfavorável entre hidrocarbonetos e água, elas são chamadas de resíduos "hidrofóbicos". A associação de resíduos hidrofóbicos no interior das proteínas, predita por W. J. Kauzmann, antes das primeiras determinações de estruturas de proteínas, é uma contribuição importante para a estabilidade protéica. Este efeito é análogo à formação de gotas de óleo no molho de saladas (ver Quadro O efeito hidrofóbico).
- Forma e rigidez A forma geral de uma cadeia lateral depende de sua estrutura química e de seus graus de liberdade conformacionais internos.

Estabilidade e desnaturação de proteínas

Quais são as forças químicas que estabilizam as estruturas nativas de proteínas? Qual é o processo pelo qual, a partir de um conjunto de conformações desnaturadas, uma proteína se enovela em um único estado nativo?

Para responder a essas perguntas, os bioquímicos estudaram a desnaturação de proteínas em resposta ao calor, ou ao aumento da concentração de uréia ou hidrocloreto de guanidina (agentes desnaturantes de uso comum). Algumas medidas são *estáticas* – determinação da quantidade em equilíbrio da proteína em seu estado nativo e desnaturado, sob diferentes condições, ou o calor liberado ao longo da transição de um estado para outro. Outras são *cinéticas* – medida das taxas de enovelamento ou desenovelamento, ou a identificação de estruturas transientes que surgem ao longo do processo.

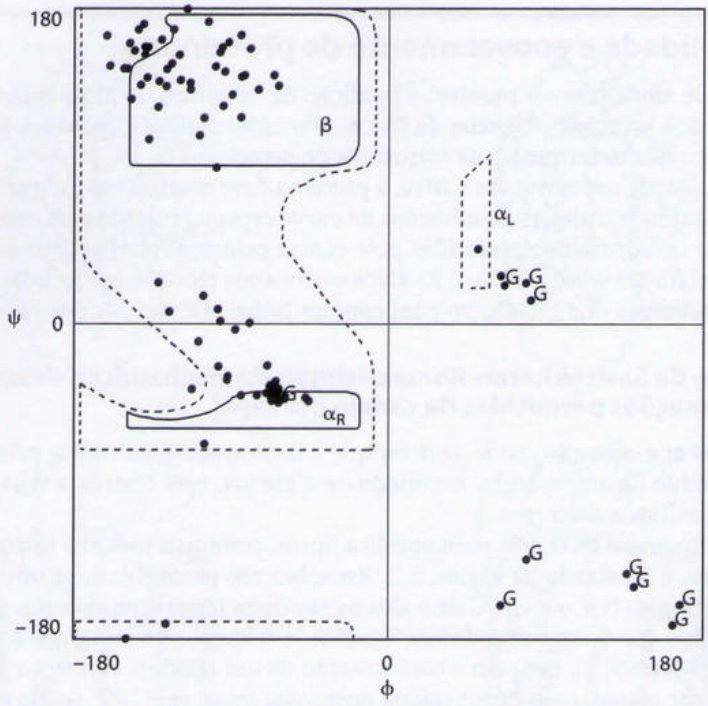


Figura 5.3 O mapa de Sasisekharan-Ramakrishnan-Ramachandran da acilfosfatase [código PDB: 2ACY]. Note a concentração de resíduos nas regiões α e β. A maioria das exceções ocorre com os resíduos de glicina (identificados por G).

de Sasisekharan-Ramakrishnan-Ramachandran – normalmente abreviado como mapa de Ramachandran (ver Figura 5.3). As linhas sólidas na figura delimitam as regiões de ϕ e ψ energeticamente preferidas; as linhas tracejadas delimitam as regiões estericamente proibidas. As conformações da maioria dos aminoácidos ficam dentro das regiões α_R ou β. A glicina tem acesso a conformações adicionais. Mais especificamente, ela pode formar hélices levógiras: α_L. A Figura 5.3 mostra a distribuição típica de conformações dos resíduos em uma estrutura protéica bem determinada. A maioria dos resíduos encontra-se dentro, ou próxima, das regiões permitidas, apesar de alguns serem forçados, pelo enovelamento, a assumirem estados energeticamente menos favoráveis.

As regiões permitidas geram conformações-padrão. Uma seqüência de resíduos consecutivos na conformação α (normalmente 6-20 em proteínas globulares em seus estados nativos) forma uma hélice α. A repetição da conformação β gera uma fita β estendida. Duas ou mais fitas β podem interagir lateralmente para formar folhas β, como na acilfosfatase (Figura 5.1). Hélices e folhas são elementos estruturais "padrão" ou "pré-fabricados" que formam os componentes das conformações da maioria das proteínas. Eles são estabilizados por interações relativamente fracas, **ligações de hidrogênio**, entre os átomos da cadeia principal. Em algumas proteínas fibrosas, praticamente todos os resíduos pertencem a um desses dois tipos de estrutura: a lã contém hélices α; a seda, folhas β. Fibrilas amilóides, formadas em algumas doenças por várias proteínas, também contêm extensas folhas β. Interações estéricas permitem que um segmento de resíduos consecutivos se localize totalmente na conformação α_R ou totalmente na conformação β, mas proíbem uma conformação de hélice seguida de uma fita e vice-versa.

Proteínas globulares típicas contêm várias regiões de hélices e/ou folhas, conectadas por *voltas*. Normalmente, as regiões terminais de uma hélice ou fita localizam-se na superfície de um domínio de uma estrutura protéica. Elas são conectadas por voltas ou alças – regiões nas quais as cadeias mudam de direção, dirigindo-se para dentro da estrutura. A maioria das voltas, mas não todas, são curtas e localizadas em uma região exposta à superfície que tende a conter resíduos carregados ou polares.

Como a cadeia principal escolhe entre as possíveis conformações permitidas? O que é único em cada proteína é a sua seqüência de cadeias laterais. Logo, as interações que envolvem as cadeias laterais devem determinar a conformação da cadeia principal.

As cadeias laterais

As cadeias laterais oferecem a versatilidade físico-química necessária para gerar todos os padrões de enovelamento diferentes. As cadeias laterais dos vinte aminoácidos variam em:

- Tamanho A menor, a glicina, consiste em apenas um átomo de hidrogênio; uma das maiores, a fenilalanina, contém um anel benzênico.
- Carga elétrica Algumas cadeias laterais possuem uma carga efetiva positiva ou negativa em pH neutro. Asp e Glu são negativamente carregadas. Lis e Arg são positivamente carregadas. (Resíduos carregados com cargas opostas podem formar um par iônico com interações atrativas chamadas de pontes salinas.)
- Polaridade Algumas cadeias laterais são polares; elas podem formar ligações de hidrogênio com outras cadeias laterais polares, ou com a cadeia principal, ou com a água. Outras cadeias laterais são eletricamente neutras. Algumas delas contêm grupos químicos relacionados com hidrocarbonetos comuns, como metano ou benzeno. Por causa da interação termodinamicamente desfavorável entre hidrocarbonetos e água, elas são chamadas de resíduos "hidrofóbicos". A associação de resíduos hidrofóbicos no interior das proteínas, predita por W. J. Kauzmann, antes das primeiras determinações de estruturas de proteínas, é uma contribuição importante para a estabilidade protéica. Este efeito é análogo à formação de gotas de óleo no molho de saladas (ver Quadro O efeito hidrofóbico).
- Forma e rigidez A forma geral de uma cadeia lateral depende de sua estrutura química e de seus graus de liberdade conformacionais internos.

Estabilidade e desnaturação de proteínas

Quais são as forças químicas que estabilizam as estruturas nativas de proteínas? Qual é o processo pelo qual, a partir de um conjunto de conformações desnaturadas, uma proteína se enovela em um único estado nativo?

Para responder a essas perguntas, os bioquímicos estudaram a desnaturação de proteínas em resposta ao calor, ou ao aumento da concentração de uréia ou hidrocloreto de guanidina (agentes desnaturantes de uso comum). Algumas medidas são *estáticas* – determinação da quantidade em equilíbrio da proteína em seu estado nativo e desnaturado, sob diferentes condições, ou o calor liberado ao longo da transição de um estado para outro. Outras são *cinéticas* – medida das taxas de enovelamento ou desenovelamento, ou a identificação de estruturas transientes que surgem ao longo do processo.

O efeito hidrofóbico

A diferença entre as diversas cadeias laterais de aminoácidos, em relação à sua preferência por ambientes aquosos ou do tipo oleoso, é um dos princípios que governam a estrutura de proteínas.

O que é o efeito hidrofóbico? É a solubilidade discreta de solutos não-polares em água resultante da estrutura microscópica da água líquida em torno desses solutos. A separação de fases em misturas óleo-água – por exemplo, no molho de saladas – é um exemplo comum. Outro exemplo são os gases (ao contrário da maioria dos sólidos) que se tornam menos solúveis em água à medida que a temperatura aumenta. Leitores que possuem chaleiras com apito devem ter escutado baixos níveis de sons antes mesmo de a água entrar em ebulição, pois o ar dissolvido sai da solução à medida que a água é aquecida.

Qual é a origem do efeito hidrofóbico? A água fria é um líquido altamente estruturado. Ela contém muitas ligações de hidrogênio que contribuem para o seu alto calor de vaporização e baixa densidade. Mas a água é ainda muito mais organizada em torno de solutos do que no líquido puro. O metano dissolvido em água – ele é apenas ligeiramente solúvel, mas solúvel o suficiente para ser estudado – é envolvido por uma "gaiola" de moléculas de água chamada de complexo clatrato. Como resultado, a dissolução de metano em água torna o solvente ainda *mais* ordenado, diminuindo a entropia. A tendência natural em direção ao estado de maior entropia inibe a dissolução de metano em água. Por isso, o metano e outros hidrocarbonetos são apenas levemente solúveis em água. As solubilidades de gases não-polares diminuem com o aquecimento do sistema – de um valor já baixo em água fria – porque, à medida que a temperatura sobe, a entropia desempenha um papel ainda mais importante na determinação do estado de equilíbrio.

O efeito hidrofóbico em soluções aquosas de solutos não-polares simples era bem conhecido pelos físico-químicos quando W. J. Kauzmann, em 1959, reconheceu a sua importância para a estrutura de proteínas.

As cadeias laterais não-polares de proteínas são similares aos solutos do tipo óleo. Suas interações com a água são desfavoráveis. Kauzmann previu que elas seriam mantidas no interior de proteínas, distantes do solvente. Esse *modelo gota-de-óleo do interior de proteínas* foi confirmado por estruturas cristalinas de proteínas globulares, resolvidas por difração de raios X. Atualmente, também reconhecemos a importância da alta densidade de empacotamento no interior de proteínas, e que é melhor tratar o interior de uma proteína enovelada mais como um cristal do que como um líquido orgânico. Contudo, o efeito hidrofóbico não perdeu a sua importância.

A cadeia principal deve atravessar os interiores da proteína, e carrega consigo os átomos polares N e O dos grupamentos peptídicos, os quais podem interagir com outros átomos polares da cadeia principal e com cadeias laterais polares, tais como treonina ou asparagina. Dessa forma, o interior não é completamente do tipo óleo. Entretanto, resíduos carregados são quase completamente excluídos dos interiores de proteínas; em casos raros, eles formam pontes salinas internas. Por outro lado, a superfície de uma proteína não é exclusivamente carregada ou polar. Em torno de metade dos resíduos na superfície de uma proteína são não-polares.

Uma mensagem muito importante é que proteínas são apenas levemente estáveis. O estado nativo de proteínas globulares é normalmente apenas 20-60 kJ mol^{-1} (5-15 kcal mol^{-1}) mais estável do que o estado desnaturado. Isso é equivalente a aproximadamente duas ligações de hidrogênio água-água.

O porquê das proteínas serem levemente estáveis não é claro. Algumas pessoas acreditam que isso facilita a reciclagem de proteínas (degradação e ressíntese de proteínas). Outras sugerem que proteínas são tão estáveis quanto elas precisam ser, então, "por que se incomodar" (menos informalmente: não há nenhuma vantagem seletiva) em otimizar ainda mais as interações estabilizantes. Sabemos, de fato, que as interações que estabilizam as proteínas nativas são capazes de produzir estruturas protéicas muito mais estáveis.

Suponha que você seja uma proteína globular em solução aquosa e que você queira atingir um estado nativo estável. Seu maior problema é a grande perda de liberdade conformacional, em relação ao conjunto de estados desnaturados, que é exigida de você quando adotar uma conformação exclusiva. Isso impõe uma grande redução da entropia que é termodinamicamente desfavorável. Uma forma de compensar isso é formar um estado globular compacto, isolando vários resíduos no interior, distantes do contato com a água. A liberação da água de interações com os átomos não-polares da proteína produz um efeito compensatório de *aumento* da entropia devido ao efeito hidrofóbico (ver Quadro).

Tudo bem, mas agora você descobre que, para formar este estado compacto, você deve isolar muitos átomos polares, incluindo mas não se limitando aos nitrogênios e oxigênios carbonílicos da cadeia principal. No estado desnaturado, esses átomos fazem ligações de hidrogênio com a água. Quando isolados no interior, seu potencial de ligação de hidrogênio deve ser satisfeito de alguma forma. (Não se esqueça: uma ou duas ligações de hidrogênio incompensadas, e você estragou tudo; o seu estado nativo seria instável.) Uma solução conveniente e comum que satisfaz o potencial de formação de ligações de hidrogênio da cadeia principal é formar hélices ou folhas.

Há um bônus: a formação de uma estrutura secundária também garante que a cadeia principal esteja em uma conformação estericamente aceitável, como delimitado pelo mapa de Sasisekharan-Ramakrishnan-Ramachandran. Resíduos em hélices α estão todos na conformação α; resíduos nas fitas de uma folha β estão todos na conformação β.

Como você decide qual região deve formar hélices ou fitas? Do ponto de vista da entalpia, hélice e folha são razoavelmente semelhantes para a maioria dos resíduos. Contudo, entropicamente, algumas cadeias laterais são estericamente mais restritas em hélices do que em fitas; estas preferem fitas. Estes efeitos influenciam a formação de estruturas secundárias. Seqüências específicas que promovem a formação de ligações de hidrogênio entre cadeia lateral e cadeia principal formam os **remates de uma hélice** (*helix caps*), determinando onde as hélices α começam e terminam.

Quão compacto deve ser o estado globular? Você poderia obter a exclusão de água do seu interior com um leve empacotamento – desde que nenhum canal tenha um raio maior do que 1,4 Å (o tamanho de uma molécula de água). Porém, quanto mais próximo conseguir arranjar seus átomos, mais vantagens você pode tirar das forças de van der Waals, forças comuns de atração entre átomos que dão à matéria a sua coesão básica. Os interiores de proteínas são densamente empacotados: o encaixe das cadeias laterais é como um grande quebra-cabeça montado. Entretanto, as peças desse quebra-cabeça (os resíduos) são

deformáveis, de modo que o processo de enovelamento é mais complicado do que o encaixe rígido das peças dos quebra-cabeças comuns.

Em resumo, você tem de encontrar uma conformação da cadeia que, simultaneamente, resolva todos os seguintes problemas:

1. Todos os resíduos devem ter uma conformação permitida estericamente. Isso se aplica tanto à cadeia principal quanto às cadeias laterais. Colisões estéricas aumentariam a energia da conformação e a tornaria instável.
2. Átomos polares isolados no interior da proteína devem fazer ligações de hidrogênio com outros átomos polares internos. Se você deixa de fazer umas poucas ligações de hidrogênio, a proteína preferirá formar o estado desnaturado de forma a permitir que esses átomos polares façam ligações de hidrogênio com o solvente.
3. Uma grande parte da superfície hidrofóbica deve ser isolada no interior da proteína, e esse interior deve ter um empacotamento suficientemente denso para garantir a estabilidade termodinâmica.

Para a maioria das proteínas, só existe uma solução para todos estes problemas, e isso define o estado nativo. Algumas proteínas mudam de conformação quando se ligam a ligantes, ou quando passam por estados metaestáveis, como parte de seus mecanismos de ação.

O fato de uma conformação de uma proteína – o estado nativo – ter uma estabilidade significativamente maior do que as outras conformações é complexo, mas não misterioso. É uma questão de otimização das interações disponíveis, e de seleção de seqüências nas quais esta conformação ótima é única e de energia significativamente menor do que as outras. Para a maioria das regiões, a estrutura local é definida pelas interações locais. Por isso, se o estado nativo não for único, deveria existir mais de uma maneira de montar um conjunto de peças deste quebra-cabeça. Dadas as restrições da cadeia, para a evolução é fácil evitar este problema.

Enovelamento de proteínas

Suponha, novamente, que você seja uma proteína, e que você esteja desnaturado. Agora que você entende como seu estado nativo é estabilizado, como você chegaria a ele? É claro que você não pode tentar todas as conformações – muitos anos atrás, C. Levinthal calculou que uma busca conformacional simples, usando números razoáveis de velocidade de rotações internas, necessitaria muito, mas muito tempo. Duas particularidades conspiram para tornar este *processo* de enovelamento das proteínas aos seus estados nativos tão misterioso e complexo.

Primeiro, é o fato de que as proteínas são apenas levemente estáveis. Isso implica que qualquer intermediário quase-estável no enovelamento da proteína teria de ser ainda menos estável, ou do contrário o processo de enovelamento ficaria preso nos intermediários. De fato, para muitas proteínas, as medidas de frações de moléculas em seus estados nativos e desnaturados, como função da temperatura ou do agente desnaturante, implica simplesmente um equilíbrio entre dois estados, nativo $\leftrightarrow$ desnaturado, no qual poucas moléculas, quase indetectáveis, são quaisquer coisas menos nativas ou desnaturadas. Isso confirma que quaisquer possíveis intermediários não podem ter mais do que uma estabilidade marginal. Mas isso torna ainda mais difícil seguir, estruturalmente, a transição para o enovelamento.

A segunda particularidade que torna o enovelamento de proteínas misterioso é que o estado desnaturado é tão heterogêneo que, na ausência de intermediários estáveis, não há nenhuma forma conveniente de se visualizar a via de enovelamento completa.

Compare o enovelamento de proteínas com outros dois tipos de formação de estrutura:

1. Na montagem de móveis do tipo "faça-você-mesmo", passa-se por uma seqüência de estados intermediários bem definidos. Primeiro, parafusa-se A em B, na conformação tipo nativa. A estrutura do fragmento A-B é determinada e estabilizada simplesmente pelas interações entre A e B. Se não fosse pela força da gravidade, um intermediário A-B seria formado. Mas as proteínas não desfrutam do luxo de formarem estados intermediários.
2. Na montagem de um arco a partir de suas diversas peças arcadas, a estrutura como um todo não é estável até a colocação da peça mestra. Somente o arco completo é independentemente estável. Não há intermediários estáveis, e a única forma de montar a estrutura é usando um andaime, o qual é subseqüentemente removido. Mas as proteínas não têm o luxo de poderem usar andaimes externos.

O que as proteínas têm de fazer é lidar com intermediários instáveis – como os móveis do tipo "faça-você-mesmo" na *presença* de força da gravidade – e terminar o processo antes de os intermediários se desfazerem, ou então continuar reformando-os e tentar continuamente.

A identificação de estruturas transientes durante o enovelamento de proteínas pode ser obtida experimentalmente por meio de medidas de troca isotópica. Prepare uma amostra de proteína desnaturada na qual todos os átomos de hidrogênio são substituídos por deutério. (É possível separar os sinais de H e D em experimentos de NMR.) Em diversos instantes durante o processo de enovelamento, em experimentos separados, exponha a amostra a um pulso de prótons. Depois que o estado nativo estiver formado, identifique onde e quando ocorreu a troca D ↔ H na estrutura. Tais estudos corroboram o modelo de que muitas proteínas se enovelam pela formação inicial de um glóbulo amorfo (*molten globule*) contendo algumas estruturas secundárias, mas sem as interações estruturais terciárias que sustentam a molécula em sua conformação final. Isso é seguido por uma condensação hierárquica para formar estruturas supersecundárias, etc., levando, por fim, à formação da estrutura nativa. Para a maioria das proteínas, não há evidências de estruturas não-nativas como intermediárias ao longo da via de enovelamento, embora estruturas não-nativas – como isômeros de prolina incorretos – possam deslocar a via, atrasando o processo de enovelamento.

A conclusão é que as estruturas de regiões locais são determinadas principalmente por interações locais, e, embora essas interações devam ser inadequadas para estabilizar regiões locais a ponto de poderem ser isoladas, elas são boas o suficiente para garantir uma via de baixa energia para a formação da estrutura.

Aplicações da hidrofobicidade

Utilizando uma **escala de hidrofobicidade** que atribui um valor para cada tipo de aminoácido, podemos calcular a variação da hidrofobicidade ao longo da seqüência de uma proteína. Isso é chamado de **perfil de hidrofobicidade**. A análise de perfis de hidrofobicidade tem sido utilizada para predizer as posições de voltas entre elementos de estrutura secundária, resíduos expostos ou isolados no interior da proteína, segmentos que atravessam membranas e regiões antigênicas.

Exemplo 5.1 O uso de perfis de hidrofobicidade na predição da posição de voltas entre hélices e fitas de uma folha β

A Figura 5.4a mostra o perfil de hidrofobicidade da lisozima da clara de ovo de galinha. Ela tem mínimos pronunciados nos seguintes resíduos: 17, 44, 70, 100 e 117. A Figura 5.4b mostra a estrutura da lisozima da clara de ovo de galinha, na qual é possível verificar a correlação entre as alças na estrutura e a posição dos mínimos no perfil de hidrofobicidade.

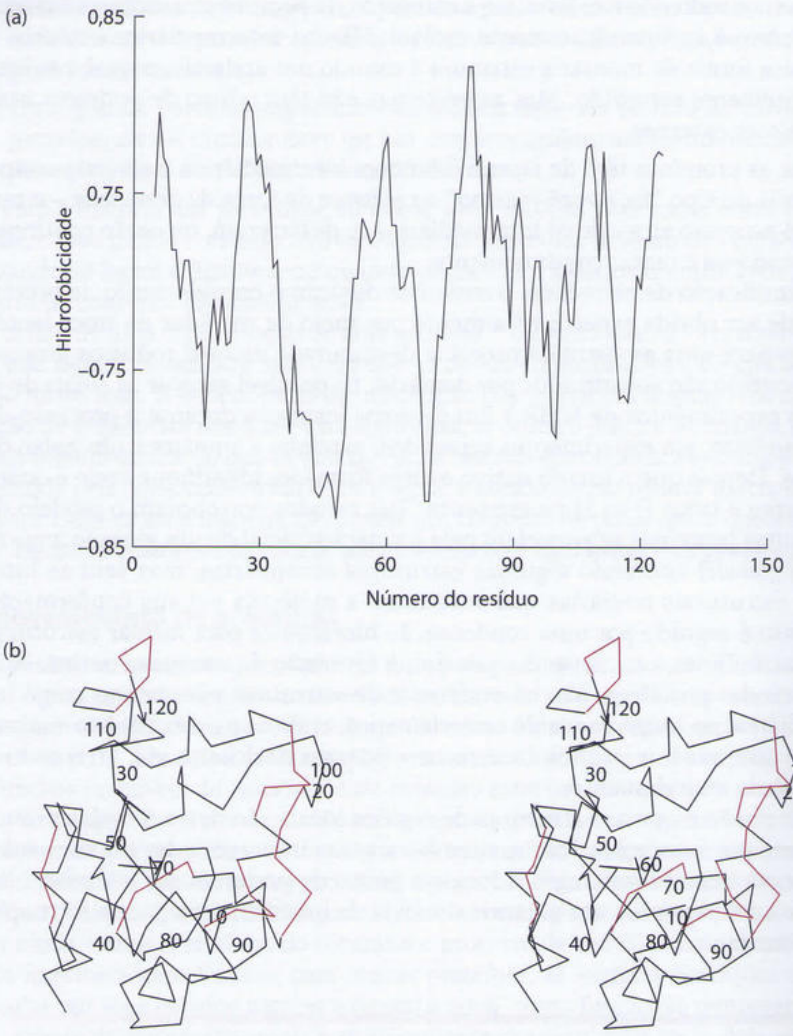


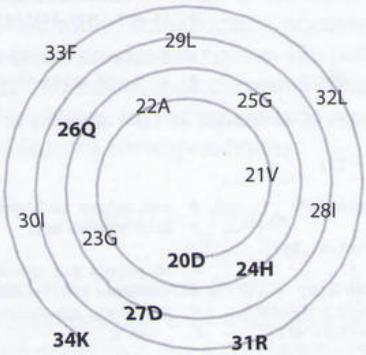
Figura 5.4 (a) Perfil de hidrofobicidade da lisozima da clara de ovo de galinha. (Produzida usando a ferramenta ProtScale em Primary Structure Analysis disponível no endereço `http://www.expasy.org`.) (b) Estrutura da lisozima da clara de ovo de galinha. As regiões correspondentes aos mínimos no gráfico de hidrofobicidade estão representadas em vermelho.

Quatro dos mínimos principais no perfil de hidrofobicidade aparecem nas ou próximos às posições das voltas. Outro mínimo ocorre em uma região exposta à superfície, mas, na estrutura, ele corresponde a uma fita de uma folha β, em vez de uma volta. Um dos mínimos está localizado em uma hélice. Da mesma forma, muitas das voltas não correspondem a mínimos pronunciados no gráfico de hidrofobicidade. Os perfis de hidrofobicidade fornecem informações úteis, mas não predizem inequivocamente todas as voltas em uma estrutura de proteína.

Exemplo 5.2 A roda helicoidal

O. B. Ptitsyn observou que hélices α em proteínas globulares geralmente possuem uma "face hidrofóbica" voltada para o interior da proteína e uma "face hidrofílica" voltada para o solvente. Cada resíduo em uma hélice α aparece em uma posição a 100° ao longo de uma circunferência a partir do seu predecessor. Por isso, para ter-se o efeito Ptitsyn, a seqüência de resíduos deveria alternar entre hidrofóbico e hidrofílico com uma periodicidade de aproximadamente quatro resíduos.

Para conferir esta relação, os resíduos podem ser projetados em um plano perpendicular ao eixo da hélice, em um diagrama denominado **roda helicoidal (*helical wheel*)**. Este exemplo mostra a seqüência de uma hélice α da mioglobina da baleia Sperm. Os resíduos carregados e polares aparecem em negrito; os outros, em caracteres comuns.



A hélice possui uma face hidrofóbica – que se direciona para o interior da estrutura – e uma face hidrofílica – que se direciona para o exterior. A partir deste padrão de hidrofobicidade, podemos prever se uma região de uma seqüência de aminoácidos poderá formar uma hélice α na estrutura de uma proteína nativa.

O próximo quadro mostra um programa PERL para desenhar rodas helicoidais.

Exemplo PERL 5.1 Um programa para desenhar rodas helicoidais

```
#!/usr/bin/perl
#helwheel.pl -- draw helical wheel
#usage: echo DVAGHGQDILIRLFKSH | helwheel.prl > output.ps
# or    echo 20DVAGHGQDILIRLFKSH | helwheel.prl > output.ps
#       the numerical prefix sets the first residue number

# The output of this program is in PostScript (TM),
#      a general-purpose graphical language

# The next section prints a header for the PostScript file

print <<EOF;
%!PS-Adobe-
%%BoundingBox: (atend)
%1 0 0 setrgbcolor
%newpath
%37.5 161 moveto 557.5 161 lineto 557.5 681 lineto 37.5 681 lineto
%closepath stroke
297.5 421. translate 2 setlinewidth 1 setlinecap
/Helvetica findfont 20 scalefont setfont 0 0 moveto
EOF

#  Define fonts to associate with each amino acid

$font{"G"} = "Helvetica";      $font{"A"} = "Helvetica";      $font{"S"} = "Helvetica";
$font{"T"} = "Helvetica";      $font{"C"} = "Helvetica";      $font{"V"} = "Helvetica";
$font{"I"} = "Helvetica";      $font{"L"} = "Helvetica";      $font{"F"} = "Helvetica";
$font{"Y"} = "Helvetica";      $font{"P"} = "Helvetica";      $font{"M"} = "Helvetica";
$font{"W"} = "Helvetica";      $font{"H"} = "Helvetica-Bold"; $font{"N"} = "Helvetica-Bold";
$font{"Q"} = "Helvetica-Bold"; $font{"D"} = "Helvetica-Bold"; $font{"E"} = "Helvetica-Bold";
$font{"K"} = "Helvetica-Bold"; $font{"R"} = "Helvetica-Bold";

$_= <>;                                  #  read line of input
chop();$_ =~ s/\s//g;                    #  remove terminal carriage return and blanks

if ($_ =~ s/^(\d+)//)                    #  if input begins with integer
     {$resno = $1;}                      #  extract it as initial residue number
else {$resno = 1}                        #  if not, set initial residue number = 1

$radius = 50;                            #  initialize values for radius,
$x = 0; $y = -50; $theta = -90;          #  x, y and angle theta

#  print light gray spiral arc as succession of line segments, 10 per residue

$npoints = 10*(length($_) - 1);

print "0.8 0.8 0.8 setrgbcolor\n";       #  set colour to light gray
print "newpath\n";                       #  draw spiral arc
printf("%8.3f %8.3f moveto\n",$x,$y);
foreach $d (1 .. $npoints) {             #  10 points per residue
    $theta += 10; $radius += 0.6;        #  increase radius and theta
    $x = $radius*cos($theta*0.01747737); #  calculate new value of x
    $y = $radius*sin($theta*0.01747737); #    and y

     printf("%8.3f %8.3f lineto\n",$x,$y);
}
print "stroke\n";

#  print residues and residue numbers

$radius = 50;                            #  reinitialize values for radius,
$x = 0; $y = -50; $theta = -90;          #  x, y and angle theta
print "0 setgray\n";                     #  set colour to black
```

→

```
foreach (split ("",$_)) {                    #  loop over characters from input line
    print "/$font{$_} findfont ";            #  set font appropriate
    print "20 scalefont setfont\n";          #  for this amino acid
    printf("%8.3f %8.3f moveto\n",$x,$y);    #  move to current point
    print " ($resno$_) stringwidth";         #  adjust position to center residue
    print " pop -0.5 mul -7 rmoveto\n";      #     identification on point on spiral
    print " ($resno$_) show\n";              #  print residue number and id
    print "% $theta $resno$_\n";
    $theta += 100; $radius += 6;             #  set new values of angle, radius
    $x = $radius*cos($theta*0.01747737);     #  compute new values of x
    $y = $radius*sin($theta*0.01747737);     #     and y
    $resno++;                                #  increase residue number
}

print "showpage\n";                          #  postscript signals to
print "%%BoundingBox:";                     #  print
$xl = 297.5 - 1.05*$radius;                  #   x
$xr = 297.5 + 1.05*$radius;                  #     and
$yb = 421.  - 1.05*$radius;                  #      y
$yt = 421.  + 1.05*$radius;                  #       limits
printf("%8.3f %8.3f %8.3f %8.3f\n",$xl,$xr,$yb,$yt);


print "showpage\n";
print "%%EOF\n";                            #  and wind up
```

Sobreposição de estruturas e alinhamentos estruturais

Alguns aspectos da análise de seqüência relacionam-se quase diretamente com a análise estrutural, alguns devem ser generalizados, e outros não têm qualquer analogia.

Como no caso de seqüências, uma questão fundamental na análise de estruturas é desenvolver e calcular uma medida de similaridade. Se duas moléculas têm estruturas idênticas ou muito semelhantes, podemos tentar sobrepô-las de modo que seus pontos correspondentes fiquem tão próximos quanto possível. Então, a distância média entre os pontos correspondentes é uma medida da similaridade estrutural. Na prática, foi convencionado representar o desvio médio quadrático (RMSD) dos átomos correspondentes.

$$\text{RMSD} = \sqrt{\sum d_i^2 / n}$$

Onde d_i é a distância entre o $i^{ésimo}$ par de átomos (um átomo de cada estrutura) após o ajuste ótimo da sobreposição, e n é o número de pontos.

Isso assume que já pré-especificamos a correspondência entre os pontos; ou seja, o alinhamento.

Se a correspondência não é conhecida, primeiro devemos determiná-la e, só então, calcular o RMSD das subestruturas alinhadas. Se cada ponto corresponde a um átomo representando os sucessivos resíduos da estrutura de uma proteína ou ácido nucléico (os átomos Cα de proteínas ou os átomos de fósforo de ácidos nucléicos), o problema é, literalmente, uma questão de alinhamento (= atribuição de correspondência entre resíduos equivalentes) (ver Quadro na página 255). Na verdade, a determinação das correspondências entre resíduos equivalentes via sobreposição espacial de duas ou mais proteínas é um método eficiente de alinhamento de seqüências. Como a estrutura tende a divergir de

forma mais conservativa do que a seqüência ao longo da evolução, o alinhamento de estruturas é um método mais eficiente do que o alinhamento par a par de seqüências para a detecção de homologia, e para o alinhamento de seqüências e medida de similaridade estrutural entre proteínas distantemente relacionadas (ver Quadro na página 255).

Exemplo 5.3 Alinhamento estrutural da quimotripsina γ e da toxina epidermolítica A de *Staphylococcus aureus*

A quimotripsina e a toxina epidermolítica A de *S. aureus* são membros da família quimotripsina de proteinases. A Figura 5.5 mostra a sobreposição estrutural das entradas PDB 8GCH (quimotripsina γ) (preto) e 1AGJ (toxina epidermolítica A de *S. aureus*) (vermelho). As moléculas compartilham o padrão de enovelamento comum da família quimotripsina das serina proteinases e a tríade catalítica Ser-His-Asp (linhas mais grossas). Segue um alinhamento das seqüências derivado da sobreposição estrutural:

```
8gch CGVPAIQPVLIVNG-----------------------------EEAVP--GS----WPWQVSLQ-DKTG
1agj -------------EVSAEEIKKHEEKWNKYYGVNAFNLPKELFSKVDEKDR-QKYPYNTIGNVFVK-G-

8gch FH--FCGGSLINE-NWVVTAAHC-GV-T---T-SDVVVAGEFDQG---SSSEKI--QKLKIAKVFK-NS-
1agj --QTSATGVLIG-KNTVLTNRHIAK-FANGDPSKVSFRPSI-NTDDNGNT-E-TPYGEYEVKEILQEP-F

8gch KYNSLTINNDITLLKLST-----AAS--FSQTVSAVCLPSASD--DFAAGTTCVTTGWG-LTRYNTPD-R
1agj GAG-----VDLALIRLKPDQNGVSL-GDK---ISPAKIGT---SNDLKDGDKLELIGYPFDH----KVNQ

9gch LQQASLPLL-SNTNCKKYWGTKIKDAM--ICAGASGV-SSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTC
1agj MHRSEIELTTLS---------------RGLRYY----GFTVPGNSGSGIFNSN---GELVGIHSSK----

8gch STST---------PGVYARVTA-LVNWVQQTLAAN-
1agj ----VSHLDREHQINYGVGIGNYVKRIINEKN---E
```

A similaridade entre estas duas seqüências está bem dentro da "zona de incerteza". Ela não poderia ser derivada corretamente a partir apenas do alinhamento par a par padrão das duas seqüências.

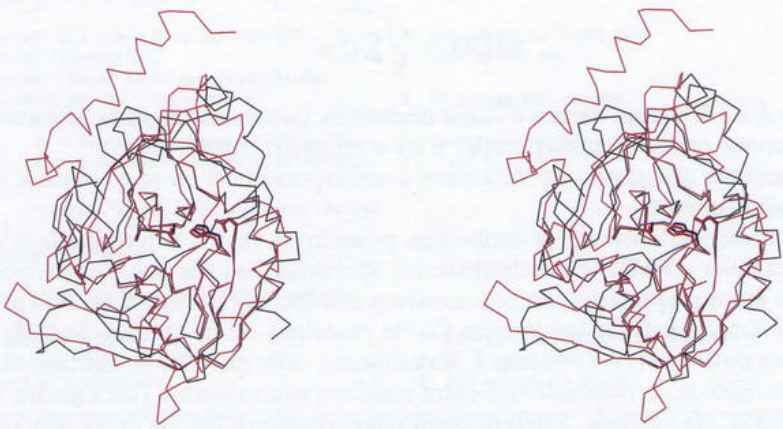


Figura 5.5 Sobreposição da estrutura da quimotripsina γ [código PDB: 8GCH] (preto) e da toxina epidermolítica A de *S. aureus* [código PDB: 1AGJ] (vermelho). As cadeias laterais da tríade catalítica estão destacadas. Observe que a região em torno do sítio ativo é a região mais bem conservada da proteína.

Determinação de similaridade e alinhamento na química computacional

1. Similaridade de dois conjuntos de átomos de correspondência conhecida:

$$p_i \leftrightarrow q_i,\ i = 1, \ldots N.$$

Para seqüências, o análogo é a distância de Hamming: somente para resíduos não correspondentes (malpareados).

2. Similaridade de dois conjuntos de átomos de correspondência desconhecida, mas cuja estrutura molecular – especificamente, a ordem linear dos resíduos – restringe as possibilidades. No caso de proteínas ou ácidos nucléicos, estamos limitados às correspondências nas quais mantemos a ordem ao longo da cadeia:

$$p_{i(k)} \leftrightarrow q_{j(k)},\ k=1, \ldots K \leq N,M$$

com a seguinte restrição: $k_1 > k_2 \Rightarrow i(k_1) > i(k_2)$ e $j(k_1) > j(k_2)$. Isso pode ser pensado como análogo à distância de Levenshtein, ou ao alinhamento de seqüência com lacunas. O resultado de um cálculo como este é o alinhamento parcial ou total de todas as seqüências.

3. Similaridade entre dois conjuntos de átomos de correspondência desconhecida, sem restrições na correspondência:

$$p_{i(k)} \leftrightarrow q_{j(k)}$$

Este problema surge nos seguintes casos importantes: suponha que duas (ou mais) moléculas possuam efeitos biológicos semelhantes, tal como uma mesma atividade farmacológica. Muitas vezes ocorre que essas estruturas compartilham uma distribuição comum de um subconjunto relativamente pequeno de seus átomos, que são os responsáveis pela atividade biológica. Esses átomos são chamados de farmacóforo. O problema é identificá-los: para isso, é necessário que se ache, dentre as duas ou mais moléculas, um número máximo de subconjuntos de átomos que possuem uma estrutura similar (ver Estudo de Caso 5.1).

DALI (Distance-matrix ALIgnment)

À medida que as proteínas evoluem, suas estruturas mudam. Entre os pequenos detalhes que a evolução tende a conservar fortemente estão os padrões de contatos entre resíduos. Isto é, se dois resíduos estão em contato em uma proteína, os resíduos alinhados com estes dois em uma proteína relacionada tendem a estar em contato também. Isso é verdade até mesmo para homólogos muito distantes, e mesmo se houver mudança no tamanho dos resíduos envolvidos. Mutações que alteram o tamanho dos resíduos empacotados no interior das proteínas causam ajustes no empacotamento das hélices e folhas em relação umas às outras.

L. Holm e C. Sander aplicaram estas observações ao problema do alinhamento estrutural de proteínas. Se o padrão de contato entre resíduos é preservado em proteínas distantemente relacionadas, então deveria ser possível *identificar* proteínas distantemente relacionadas detectando-se padrões de contatos conservados.

Computacionalmente, constroem-se matrizes de padrões de contato resíduo-resíduo para duas proteínas (isso é bem simples), e então se procuram as sub-

Figura 5.6 As regiões de enovelamento comum, determinadas pelo programa DALI de L. Holm e C. Sander, das proteínas do tipo *TIM-barrel* adenosina deaminase de camundongo [código PDB: 1FKX] (preto) e fosfotriesterase de *Pseudomonas diminuta* [código PDB: 1PTA] (vermelho). No alinhamento mostrado nesta figura, as seqüências possuem apenas 13% de resíduos idênticos – mais próximos da zona de total incerteza do que da zona de incerteza, em termos de identificação de homologia apenas pelo alinhamento padrão par a par de seqüências.

matrizes com os melhores pareamentos entre resíduos (isso é difícil). Utilizando aproximações cuidadosamente selecionadas, Holm e Sander escreveram um programa eficiente chamado DALI (de Distance-matrix ALIgnment) que é, atualmente, de uso comum na identificação de proteínas com padrões de enovelamento similares aos de uma proteína de interesse. O programa executa de forma rápida o suficiente para se fazer buscas rotineiras em todo o Protein Data Bank por estruturas similares a uma estrutura recém-determinada, e até mesmo classificar os domínios estruturais da proteína a partir de comparações todos-contra-todos. Holm e Sander encontraram várias similaridades inesperadas, não detectáveis por meio do alinhamento par a par de seqüências.

Como um exemplo do "alcance" do programa DALI no reconhecimento de similaridades estruturais bem distantes está a sua identificação da relação entre a adenosina deaminase de camundongo, a urease de *Klebsiella aerogenes* e a fosfotriesterase de *Pseudomonas diminuta* (ver Figura 5.6).

O programa DALI está disponível na *web*. Você pode submeter as coordenadas ao sítio `http://www2.ebi.ac.uk/dali/` e receber o conjunto de estruturas similares e seus alinhamentos com a proteína utilizada na consulta.

Evolução de estruturas de proteínas

Dentro das mais de 43.000 estruturas de proteínas atualmente conhecidas estão várias famílias nas quais as moléculas mantêm o mesmo padrão básico de enovelamento sobre uma gama de similaridade de seqüência, desde aquelas com quase 100% de identidade total até aquelas com conservação bem abaixo de 20%. As serina proteinases (quimotripsina γ e toxina epidermolítica A de *S. aureus*, Figura 5.5) e a família da adenosina deaminase-fosfotriesterase (Figura 5.6) são exemplos.

A resposta geral à mutação é a mudança estrutural. Uma característica de sistemas biológicos é que objetos que observamos possuir uma determinada forma surgiram, a partir da evolução, de outros objetos relacionados, de forma similar,

mas não idêntica. Eles devem, portanto, ser robustos para terem a liberdade de tolerar alguma variação. Podemos tirar vantagem desta robustez em nossas análises: pela identificação e por meio de comparações de objetos relacionados, podemos determinar as características conservadas e variadas, e então distinguir aquela que é essencial para a estrutura e função (e, portanto, conservada) daquela que resiste a mudanças (e, portanto, disponível para variar).

Variações naturais em famílias de proteínas homólogas que conservam a mesma função revelam como as estruturas acomodam mudanças na seqüência de aminoácidos. Os resíduos da superfície não envolvidos na função são, normalmente, livres para mutar. Alças na superfície podem, freqüentemente, acomodar mudanças pelo rearranjo conformacional local. Mutações que alteram o volume de resíduos internos geralmente não modificam a conformação de hélices e folhas individuais, mas produzem distorções em seu arranjo espacial. A natureza das forças que estabilizam as estruturas de proteínas estabelece limitações gerais nestas mudanças conformacionais; as limitações específicas derivadas da função variam de caso a caso.

As famílias de proteínas relacionadas tendem a manter um padrão de enovelamento comum. Entretanto, embora este padrão geral seja preservado, existem distorções que aumentam à medida que as seqüências de aminoácidos divergem. Essas distorções não são uniformemente distribuídas ao longo da estrutura. Normalmente, um grande *núcleo* central da estrutura retém, qualitativamente, o mesmo enovelamento, enquanto outras partes da estrutura mudam de conformação mais radicalmente. Considere as letras B e R. Como estruturas, elas têm um núcleo comum, que corresponde à letra P. Além desse núcleo comum, elas diferem na parte de baixo: à direita, a letra B tem uma volta, e a letra R tem um traço em diagonal.

Estudos sistemáticos das diferenças estruturais entre pares de proteínas relacionadas definiram uma relação quantitativa entre a divergência das seqüências de aminoácidos da região central em uma família de proteínas e a divergência de estrutura. À medida que a seqüência diverge, ocorre um aumento progressivo na distorção da conformação da cadeia principal, e a fração dos resíduos no núcleo central normalmente diminui. Até que a fração de resíduos idênticos na seqüência caia para abaixo de aproximadamente 40-50%, estes efeitos são relativamente modestos. Toda a estrutura praticamente permanece com o mesmo núcleo central, e o deslocamento dos átomos da cadeia principal, em média, não é superior a 1,0 Å. Com o aumento de divergência nas seqüências, algumas regiões se reenovelam completamente, reduzindo o tamanho do núcleo central e aumentando a magnitude das distorções dos resíduos que permanecem no núcleo.

Uma correlação entre divergência de seqüência e estrutura se aplica a todas as famílias de proteínas. A Figura 5.7a mostra as mudanças no núcleo central, expressas como o RMSD dos átomos da cadeia principal, após otimização da sobreposição, em relação à divergência de seqüência: a porcentagem de aminoácidos conservados no núcleo após alinhamento ótimo. Os pontos correspondem a pares de proteínas homólogas de muitas famílias de proteínas relacionadas. (Aquelas com 100% de identidade dos resíduos são proteínas cuja estrutura foi determinada em dois ou mais ambientes cristalinos, e os desvios mostram que as forças de empacotamento do cristal – e, em uma extensão menor, o solvente e a temperatura – podem modificar levemente a conformação das proteínas.) A Figura 5.7b mostra as mudanças na fração de resíduos no núcleo como função da divergência de seqüência. A fração de resíduos nos núcleos de proteínas distantemente relacionadas pode variar bastante: em alguns casos, a fração de resíduos no núcleo permanece alta, em outros, pode diminuir para menos de 50% da estrutura.

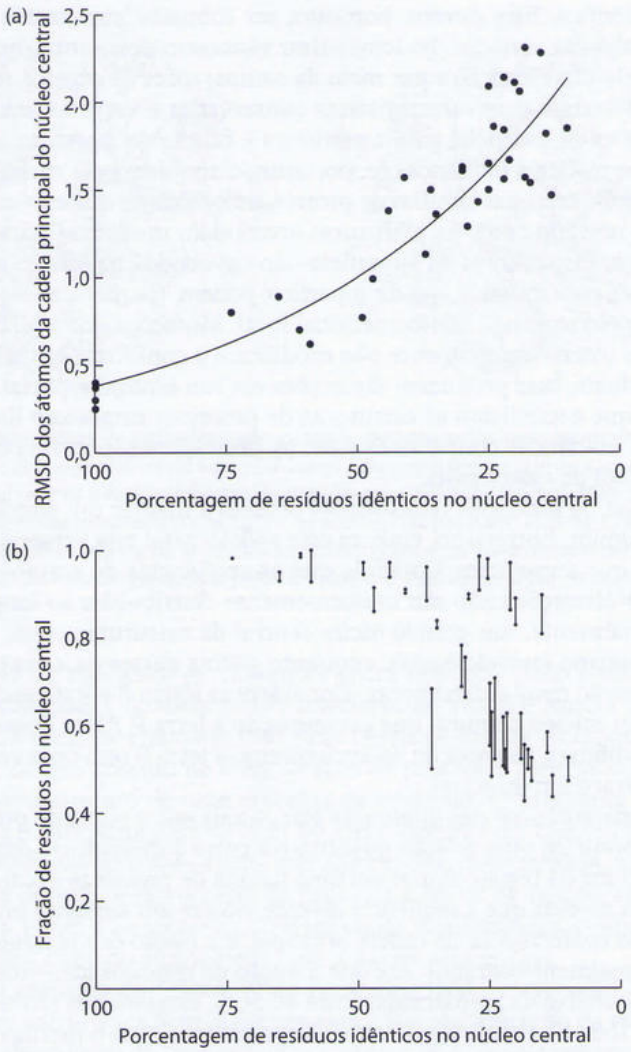


Figura 5.7 Relações entre divergência na seqüência de aminoácidos e estrutura tridimensional do núcleo central durante a evolução de proteínas. (a) Variação do RMSD do núcleo central em relação à porcentagem de resíduos idênticos no núcleo. (b) Variação do tamanho do núcleo central em relação à porcentagem de resíduos idênticos no núcleo. Esta figura mostra os resultados calculados para 32 pares de proteínas homólogas, de diversos tipos estruturais. (Adaptada de Chothia, C. & Lesk, A. M. (1986), Relationship between the divergence of sequence and structure in proteins, *The EMBO Journal*, **5**, 823-826.)

Classificação de estruturas de proteínas

A capacidade de medir quantitativamente as diferenças estruturais entre proteínas nos permite agrupá-las e classificá-las de acordo com seus padrões de enovelamento. A organização de estruturas de proteínas como função do padrão de enovelamento introduz uma organização lógica, e muito útil, das entradas no PDB. Ela fornece o embasamento para a recuperação de informações com base no tipo de estrutura. Vários bancos de dados derivados do PDB são construídos

com base na classificação de estruturas de proteínas. Eles oferecem características úteis para explorar o universo estrutural das proteínas, incluindo busca por palavras-chave ou seqüência, navegação entre estruturas similares nos vários níveis da classificação hierárquica, visualização de estruturas, pesquisa no banco de dados por estruturas similares a uma nova estrutura e conexões para outros sítios. Estes bancos de dados incluem o SCOP (Structural Classification of Proteins), CATH (Class, Architecture, Topology, Homologous superfamily), FSSP/DDD (Fold classification based on Structure-Structure alignment of Proteins/Dali Domain Dictionary), e CE (The Combinatorial Extension Method, de I. N. Shindyalov e P. Bourne).

SCOP

O SCOP, de A. G. Murzin, L. Lo Conte, B. G. Ailey, S. E. Brenner, T. J. P. Hubbard e C. Chothia, organiza hierarquicamente a estrutura de proteínas de acordo com a origem evolucionária e a similaridade estrutural. No nível hierárquico mais baixo do SCOP estão os **domínios** individuais (ver página 62), extraídos das entradas do PDB. Conjuntos de domínios são agrupados em **famílias** de homólogos, para as quais similaridades estruturais, de seqüência e, algumas vezes, de função, implicam uma origem evolucionária comum. Famílias contendo proteínas com estruturas e funções similares, mas cuja evidência de relação evolucionária é sugestiva, porém não conclusiva, formam as **superfamílias**. As superfamílias que compartilham um padrão de enovelamento (ou topologia) comum, para pelo menos uma grande parte do núcleo central da estrutura, são agrupadas como **topologias (*folds*)**. Finalmente, cada grupo de topologia reside em uma das **classes** gerais. As classes principais do SCOP são α, β, $\alpha + \beta$, α/β, e uma mista chamada de "proteínas pequenas" (*small proteins*), que, geralmente, possuem poucas estruturas secundárias e são mantidas por pontes dissulfeto ou ligantes.

O quadro a seguir mostra a classificação do SCOP para a flavodoxina do *Clostridium beijerinckii* (Gravura V). Para ilustrações do grau de similaridades de proteínas agrupadas em diferentes níveis hierárquicos e discussão de outros esquemas de classificação, ver o livro *Introduction to Protein Architecture: The Structural Biology of Proteins*, Capítulo 4.

Classificação do SCOP para a flavodoxina do *Clostridium beijerinckii*

1. **Raiz:** SCOP
2. **Classe:** proteínas α e β (α/β)
   Principalmente folhas β paralelas (unidades β-α-β)
3. **Padrão de enovelamento ou topologia:** tipo flavodoxina (*flavodoxin-like*)
   Três camadas $\alpha/\beta/\alpha$; folhas β paralelas de cinco fitas, ordem 21.345
4. **Superfamília:** flavoproteínas
5. **Família:** relacionada à flavodoxina que se liga a FMN
6. **Proteína:** flavodoxina
7. **Espécie:** *Clostridium beijerinckii*

A versão do SCOP de janeiro de 2004 continha 13.220 entradas do PDB, divididas em 31.474 domínios. A distribuição de entradas nos diferentes níveis hierárquicos é:

Classe	Número de famílias	superfamílias	topologias
Proteínas α	337	224	138
Proteínas β	276	171	93
Proteínas α/β	374	167	97
Proteínas α + β	391	263	184
Proteínas multidomínio	35	28	28
Proteínas de membrana e de superfície celular	28	17	11
Proteínas pequenas	116	77	54
Total	1.557	947	605

Vários outros sítios da *web* que oferecem classificações de estruturas de proteínas estão indexados em `http://www.bioscience.org/urllists/protdb.htm`.

Predição e modelagem de estrutura de proteínas

A observação de que cada proteína se enovela espontaneamente em uma conformação nativa tridimensional única implica que a natureza possui um algoritmo para predizer a estrutura de proteínas a partir da seqüência de aminoácidos. Algumas tentativas de se entender esse algoritmo são baseadas somente nos princípios gerais da física; outras apelam para as seqüências de aminoácidos e estruturas de proteínas conhecidas. Uma prova de nossa compreensão deste fenômeno seria a habilidade de reproduzir o algoritmo em um programa de computador que pudesse predizer a estrutura de uma proteína a partir apenas da sua seqüência de aminoácidos.

A maioria das tentativas de predição de estrutura de proteínas, a partir apenas de princípios básicos da física, tenta reproduzir as interações interatômicas em proteínas a fim de definir uma energia mensurável associada a cada conformação. Computacionalmente, o problema da predição de estrutura de proteínas torna-se, então, uma questão de se encontrar o mínimo global desta função de energia conformacional. Até agora, esta abordagem não obteve êxito, em parte devido à inadequação da função de energia e em parte porque os algoritmos de minimização tendem a ficar presos nos mínimos locais de energia.

Outras abordagens por primeiros princípios para a predição de estruturas são baseadas na tentativa de simplificar o problema, de modo a capturar, de alguma forma, os seus elementos essenciais.

Uma alternativa a estes métodos por primeiros princípios são as abordagens baseadas na pesquisa de dados sobre a estrutura de uma seqüência-alvo pela busca de similaridades com estruturas conhecidas. Esses métodos empíricos ou "baseados no conhecimento" estão se tornando bastante eficientes.

Estamos nos aproximando cada vez mais de saturar todos os tipos de enovelamentos (topologias, *folds*) possíveis com as estruturas já conhecidas. Este é o objetivo inicial dos projetos de genômica estrutural (ver Quadro). Uma vez que tenhamos um conjunto completo de tipos de topologias e seqüências, e métodos eficientes para relacioná-los, os métodos empíricos nos permitirão fornecer solu-

Genômica estrutural

Em analogia aos projetos de seqüenciamento completo de genomas, a genômica estrutural tem o compromisso de disponibilizar as estruturas do conjunto completo de proteínas. Experimentos de raios X e de NMR resolverão a estrutura de um "conjunto considerável" de proteínas, de forma que todas as proteínas fiquem dentro do limite para serem modeladas a partir de uma ou mais estruturas experimentais já conhecidas. Mais do que os projetos de seqüenciamento de genomas, os projetos de genômica estrutural combinam os resultados de diferentes organismos. O proteoma humano, assim como as proteínas exclusivas de microrganismos infecciosos, são, é claro, de interesse especial.

Os objetivos da genômica estrutural têm sido alcançados, em parte, por causa dos avanços nas técnicas experimentais, que tornaram possível a determinação de estruturas em grande escala, e, também, devido ao nosso melhor entendimento sobre estruturas de proteínas, que define objetivos gerais moderados para o trabalho experimental e sugere alvos específicos.

A teoria e a prática de modelagem por homologia sugerem que é necessária uma identidade de seqüência de pelo menos 30% entre a seqüência-alvo e alguma estrutura experimental. Isso significa que determinações de estrutura experimentais serão necessárias para um exemplar de cada família de seqüências, incluindo muitas que possuem o mesmo padrão básico de enovelamento. Os experimentos deverão fornecer as estruturas de em torno de 10.000 domínios. Em 2004, aproximadamente 5.000 estruturas foram depositadas no PDB, de modo que a taxa de produção de estruturas não está tão longe da necessária.

Métodos de bioinformática podem auxiliar na seleção de alvos para a determinação experimental de estruturas que ofereçam o melhor custo-benefício em termos de informações úteis. Os objetivos da seleção de alvos incluem:

- eliminação de alvos redundantes – proteínas muito similares a estruturas conhecidas.
- identificação de seqüências cuja similaridade com proteínas de estrutura conhecida não pode ser detectada.
- identificação de seqüências com similaridade apenas com proteínas de função desconhecida, ou
- proteínas de estrutura desconhecida com funções "interessantes"; por exemplo, proteínas humanas implicadas em doenças, ou proteínas bacterianas implicadas na resistência a antibióticos.
- proteínas com propriedades favoráveis para a determinação de estrutura – como ser solúvel, conter metionina (o que facilita a resolução do problema da fase na cristalografia por difração de raios X).

As ferramentas para se fazer modelagem já são conhecidas e funcionam. MODBASE (`http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi`) e 3DCrunch (`http://swissmodel.expasy.org/SM_3DCrunch.html`) colecionam modelos de proteínas de seqüências conhecidas derivados por homologia.

Os projetos de genômica estrutural são financiados por iniciativas em grande escala do US National Institutes of Health e da indústria privada.

ções práticas para muitos problemas. Quais serão os efeitos disso na predição da estrutura de proteínas por primeiros princípios? Os desafios intelectuais para este problema ainda estarão lá. Afinal, a natureza enovela uma proteína sem a ajuda ou a pesquisa em um banco de dados. Mas é improvável, uma vez que uma solução prática tenha sido encontrada, que este problema continue despertando interesse na mesma intensidade de antes e mantenha o mesmo nível de investimento.

Entretanto, existe um paradoxo: os métodos que estão sendo desenvolvidos para a identificação de padrões de enovelamento em seqüências representam mais do que exercícios de ajuste de parâmetros em funções de classificação. Eles são experimentos que exploram e expõem as características essenciais de seqüências de aminoácidos que determinam as estruturas de proteínas. Quando esses métodos obtiverem sucesso, teremos uma base muito mais sólida para entender as relações seqüência-estrutura do que temos hoje. Pode ser que uma compreensão baseada em conhecimentos (*a posteriori*) nos forneça dados que tornem possíveis as predições por primeiros princípios (*a priori*).

Os métodos de predição de estrutura de proteína a partir da seqüência de aminoácidos incluem:

- **Tentativas de predição da estrutura secundária** sem tentar agregar estas regiões tridimensionalmente. Os resultados são listas de regiões da seqüência que formariam hélices α e regiões que formariam fitas de folhas β.
- **Modelagem por homologia**: predição da estrutura tridimensional de uma proteína a partir de estruturas conhecidas de uma ou mais proteínas relacionadas. Os resultados são um conjunto completo de coordenadas para as cadeias principal e laterais, com a intenção de se propor um modelo estrutural de alta qualidade, comparável a, pelo menos, uma estrutura experimental de baixa resolução.
- **Reconhecimento de padrão de enovelamento ou topologia (*fold recognition*)**: dada uma biblioteca de estruturas conhecidas, determine qual delas compartilha um padrão de enovelamento com uma proteína de interesse de seqüência conhecida, mas estrutura desconhecida. Se o padrão de enovelamento da proteína-alvo não ocorre na biblioteca, tal método deveria ser capaz de fazer esta identificação. Os resultados são uma indicação de uma estrutura conhecida que tenha a mesma topologia (padrão de enovelamento) que a proteína-alvo, ou a conclusão de que nenhuma proteína da biblioteca possui a mesma topologia que a proteína-alvo.
- **Predição de novas topologias**, tanto por métodos por primeiros princípios (*a priori*) quanto por métodos baseados no conhecimento (*knowledge-based ou a posteriori*). Os resultados são um conjunto completo de coordenadas, pelo menos da cadeia principal, e algumas vezes também das cadeias laterais. O modelo deve apresentar o padrão de enovelamento correto, mas não necessariamente seria comparável em qualidade a uma estrutura experimental. D. Jones comparou a diferença entre o reconhecimento de topologia e a modelagem por primeiros princípios com a diferença entre questões de múltipla escolha e questões subjetivas em um exame.

Critical Assessment of Structure Prediction (CASP)

Os projetos CASP foram rapidamente introduzidos no Capítulo 1. O CASP organiza desafios às cegas de predição de estruturas de proteínas, nos quais os cristalógrafos ou espectroscopistas de NMR tornam pública a seqüência de aminoácidos da proteína que estão investigando e concordam em manter segredo sobre sua estrutura experimental até que os preditores (os pesquisadores ou

grupos de pesquisas que realizam as predições) submetam os seus modelos. O CASP ocorre a cada dois anos. No final do ano, um encontro de gala reúne todos os participantes para discutir os resultados encontrados e avaliar o progresso dos métodos de predição de estrutura de proteínas.

As predições do CASP tradicionalmente se distribuem em três categorias principais: (1) modelagem comparativa – efetivamente modelagem por homologia, (2) reconhecimento de topologia, e (3) modelagem de novas topologias, também conhecida como predição *ab initio* ou *de novo*:

Categoria do CASP	Natureza do alvo
Modelagem comparativa	Homólogos próximos com estrutura conhecida estão disponíveis; métodos de modelagem por homologia são empregados.
Reconhecimento de topologia	Estruturas com enovelamentos (topologias) similares estão disponíveis, mas não possuem uma relação próxima o suficiente para a modelagem por homologia; o desafio está na identificação de estruturas com topologias similares.
Novas topologias	Não existem estruturas com um padrão de enovelamento similar conhecido; há necessidade ou de um método por primeiros princípios (*a priori*) ou de um método baseado no conhecimento que combine as características de várias estruturas conhecidas.

Os assessores, um de cada categoria, comparam as estruturas preditas com a estrutura experimental e julgam as predições. Os palestrantes do encontro de final de ano incluem os organizadores, os assessores e preditores selecionados, incluindo aqueles que obtiveram sucesso ou que têm um método novo e interessante para apresentar.

O último projeto CASP aconteceu em 2004.* Diferenças das avaliações de desafios passados incluem: (1) a predição de estrutura secundária não é mais analisada separadamente, e (2) uma nova categoria, a predição de função, foi introduzida. Havia 87 alvos. Em todas as categorias, 201 grupos de preditores submeteram um total de 28.965 modelos. Isso era aproximadamente igual ao número de entradas no PDB naquela época!

Muitas predições são preparadas por grupos de pesquisadores que investigam os resultados gerados por seus programas de computador, os selecionam e editam antes de submetê-los. Além disso, as seqüências-alvo são submetidas a servidores da *web* que retornam predições sem interferência humana. O projeto CAFASP: *Critical Assessment of Fully Automated Structure Prediction* monitora a qualidade dessas predições. Assim, é possível determinar-se o quanto um método de sucesso pode ser totalmente automatizado. Portanto, o CASP compreende três desafios:

Humanos contra proteína	CASP
Computadores contra proteína	CAFASP
Humanos contra computadores	CASP vs. CAFASP

Um projeto independente de teste às cegas também avalia métodos de predição de interações proteína-proteína, ou "docagem". Ele se chama CAPRI – *Critical Assessment of PRedicted Interactions*. Tanto o CASP quanto o CAPRI tiveram seus encontros em dezembro de 2004.**

As predições de estrutura do sexto projeto CASP mostraram uma evolução nos métodos de predição. Em sua maioria, o progresso foi discreto, em vez de espetacular, com uma exceção notável: o grupo de David Baker predisse e redefi-

* N. de T. O último CASP ocorreu em 2006!

** N. de T. Os experimentos do terceiro CAPRI ocorreram em 2006, e o encontro, em abril de 2007!

niu a estrutura de uma pequena proteína (70 resíduos) de *Thermus thermophilus*, produzindo um modelo que desviava 1,59 Å da estrutura cristalográfica! Na verdade, melhorias nos métodos baseados em conhecimento, originalmente desenvolvidos para a predição de novos padrões de enovelamento, ameaçam superar os métodos tradicionais de reconhecimento de topologia, como o alinhavamento (*threading*), que faz uso explícito de bibliotecas de estruturas completas.

Os resultados do CAPRI mostraram que complexos entre proteínas que não estão sujeitos a mudanças conformacionais significativas podem, agora, ser corretamente preditos a partir das estruturas de seus componentes. Grandes mudanças conformacionais resultantes da formação de complexos ainda apresentam dificuldades. Entretanto, podem-se ver progressos em pelo menos um caso: na proteína trimérica do envelope do vírus TBE.

Tanto para CASP quanto para CAPRI, os melhores resultados são bem impressionantes. Um observador comentou que o estado atual da predição de estrutura é aquele em que "o fracasso não pode mais ser garantido". A consistência é o desafio.

Predição de estrutura secundária

Parece óbvio que (1) deveria ser mais fácil predizer a estrutura secundária do que a terciária, e (2) na predição da estrutura terciária, uma forma sensata de fazê-lo seria, primeiro, predizer as hélices e fitas de uma folha, e depois juntá-las. Não importa se estas proposições estão ou não corretas: diversas pessoas acreditam ou rebatem essas idéias. Dada a seqüência de aminoácidos de uma proteína de estrutura desconhecida, elas produzem **predições de estrutura secundária**, a atribuição das regiões de hélice ou fitas de folha na seqüência.

Para avaliar a qualidade de uma predição da estrutura secundária, classificam-se os resíduos da estrutura tridimensional experimental em três categorias [hélice = H, fita = E (estendida), e outra = –]. A porcentagem de resíduos corretamente preditos é chamada de Q3. No projeto CASP de 2000, o servidor PROF de B. Rost conseguiu uma boa predição de um domínio da proteína de reparo de malpareamento MutS de *Thermus aquaticus*. O valor de Q3 na predição de Rost foi de 81%:

```
                          10        20        30        40        50
                           |         |         |         |         |
Amino acid sequence ALVEDPPLKVSEGGLIREGYDPDLDALRAAHREGVAYFLELEERERERTG
Prediction          HH------------EEE------HHHHHHHHHH-HHHHHHHHHHHHHH-
Experiment          -E------------E-----HHHHHHHHHHHHHHHHHHHHHHHHHHHH-

                          60        70        80        90       100
                           |         |         |         |         |
Amino acid sequence IPTLKVGYNAVFGYYLEVTRPYYERVPKEYRPVQTLKDRQRYTLPEMKEK
Prediction          --EEEEEEEEEEEEEEE----------EEEEEEEE--EEEE-HHHHHH
Experiment          ----EEEEE---EEEEEEEHHHHHH-----EEEEE---EEEEE-HHHHHH

                         110       120
                           |         |
Amino acid sequence EREVYRLEALIRRREEEVFLEVRERAKRQ
Prediction          HHHHHHHHHHHHHHHHHHHHHHHHHHH-
Experiment          HHHHHHHHHHHHHHHHHHHHHHHHHHH--
```

A Figura 5.8 mostra a estrutura experimental, com a estrutura secundária *predita* em destaque. Exceto por uma pequena hélice 3_{10}, os elementos estruturais secundários são preditos corretamente, à exceção de algumas pequenas discrepâncias nas posições onde elas começam e terminam. (Outros esquemas de escore e classificação, que conferem segmentos que se sobrepõem, são menos

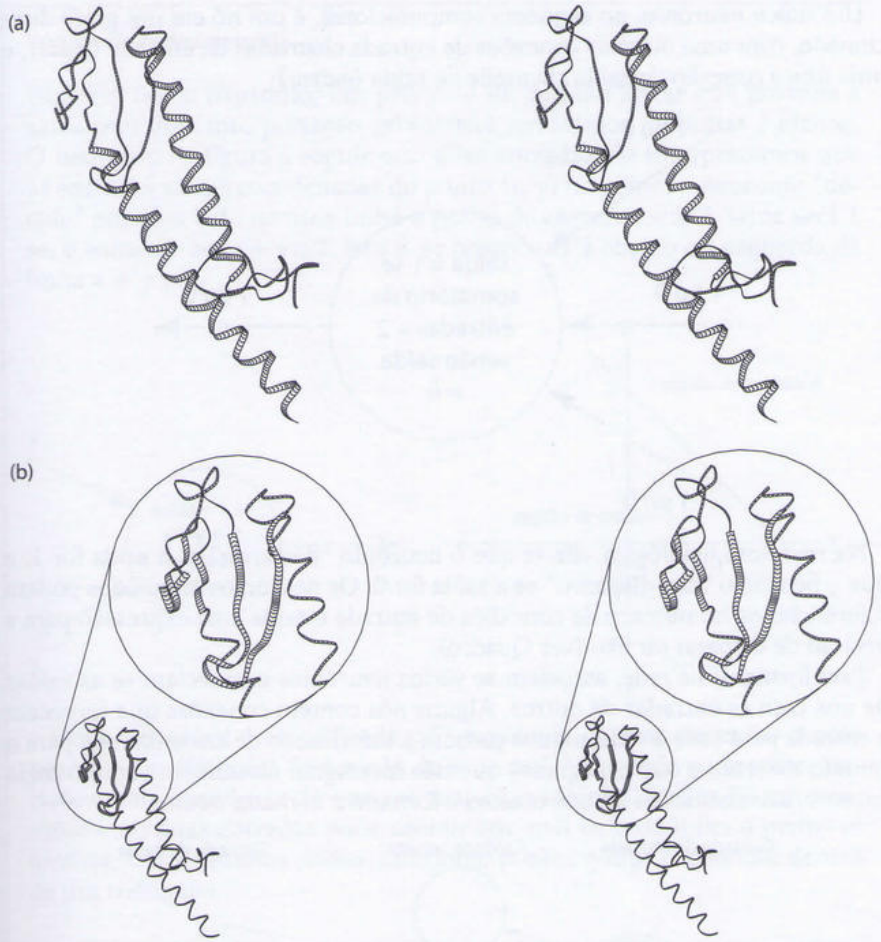


Figura 5.8 Estrutura da proteína de reparo de malpareamento MutS de *Thermus aquaticus* [código PDB: 1EWQ]. (a) As regiões preditas pelo servidor PROF de Rost como sendo helicoidais estão representadas em tiras mais largas. A predição somente perdeu uma pequena hélice 3_{10} na parte superior esquerda do desenho. (b) As regiões preditas como fitas são representadas como tiras mais largas.

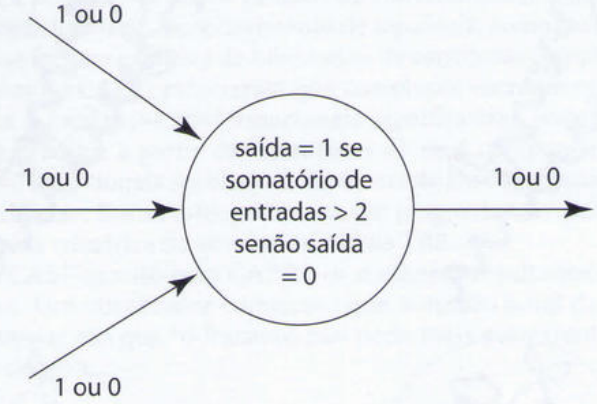
sensíveis aos efeitos nas extremidades.) A qualidade deste resultado é bastante alta, mas não é um caso excepcionalmente raro. Este alvo foi classificado como de dificuldade *média* pelos assessores do CASP4. Atualmente, o PROF está funcionando com uma precisão média de Q3 de ~77%. Outros métodos de predição de estrutura secundária também estão funcionando comparativamente bem.

Os melhores métodos de predição de estrutura secundária são baseados em **redes neurais**.

Redes neurais

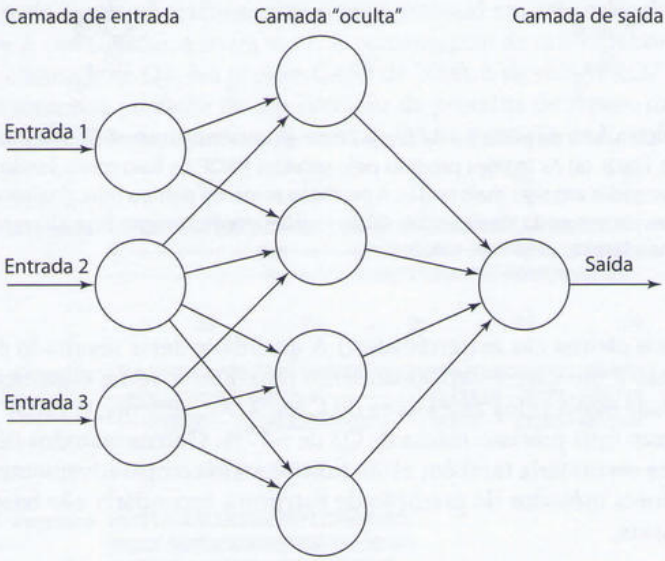
As redes neurais são uma classe de estruturas computacionais baseada na anatomia e na fisiologia de sistemas nervosos biológicos. Elas têm sido aplicadas com sucesso em uma grande variedade de problemas de reconhecimento de padrões, classificação e decisão.

Um único neurônio, no esquema computacional, é um nó em um grafo direcionado, com uma ou mais conexões de entrada chamadas de entrada (*input*), e uma única conexão de saída chamada de saída (*output*):



Na metáfora fisiológica, diz-se que o neurônio "disparou" se a saída for 1, e que o neurônio "não disparou" se a saída for 0. Os neurônios simulados podem diferenciar-se no número de conexões de entrada e saída e na expressão para a decisão de disparar ou não (ver Quadro).

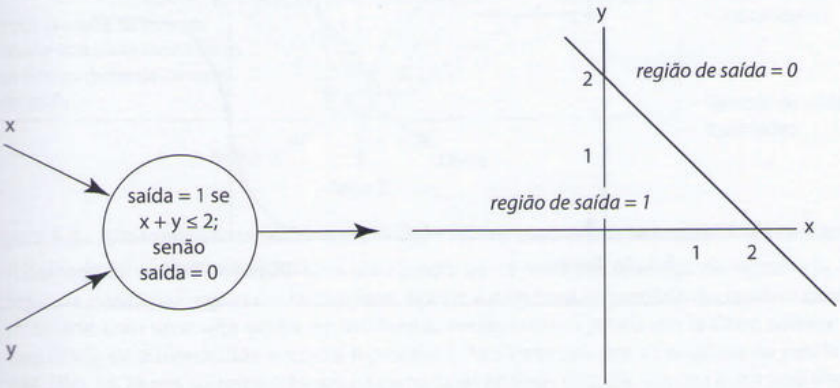
Para formar uma rede, associam-se vários neurônios e conectam-se as saídas de uns com as entradas de outros. Alguns nós contêm conexões que fornecem a entrada para toda a rede; alguns passam a informação de saída da rede para o mundo externo; e outros, aqueles que não interagem diretamente com o meio externo, são chamados de nós ocultos e formam a camada oculta.



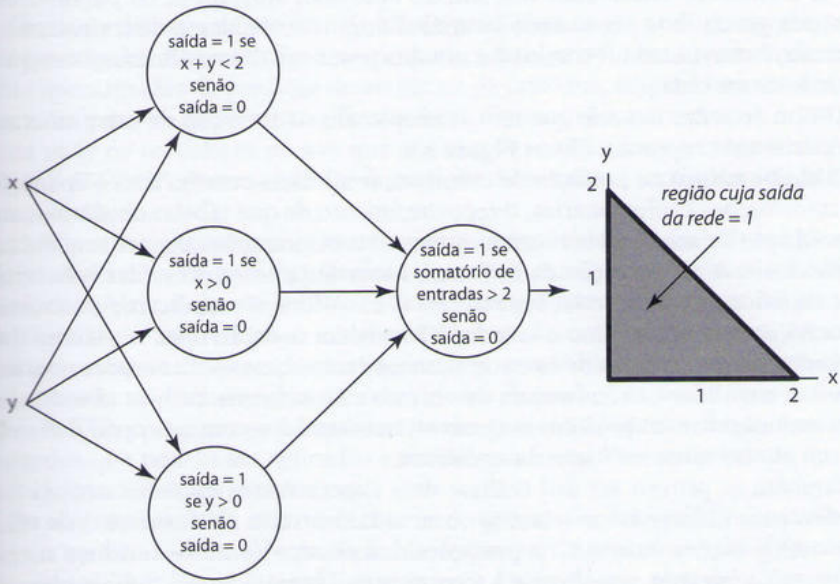
Um grau ilimitado de complexidade é possível por meio do agrupamento e da conexão de neurônios, e pela variação da magnitude das conexões. Isto é, em vez de tomar-se o somatório simples das entradas $i_1 + i_2 + i_3$, toma-se o somatório ponderado – por exemplo, $10i_1 + 5i_2 + i_3$, o que faria o neurônio mais sensível à entrada 1 e menos sensível à entrada 3. Biologicamente, isso corresponderia a modificar a intensidade das sinapses nervosas.

Lógica de redes neurais

Para um único neurônio, um processo de decisão linear que governa a saída tem uma interpretação geométrica em termos de linhas e planos. O neurônio na figura a seguir tem duas entradas. Se interpretamos que as entradas são as coordenadas do ponto (x, y) no plano, o neurônio "decide" em qual lado de uma linha o ponto de entrada está. A saída será 1 se, e somente se, $x + y \leq 2$; isto é, se o ponto está abaixo e à esquerda da linha $x + y = 2$.

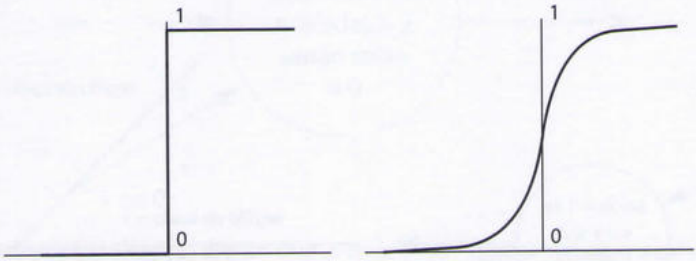


Uma *rede neural* é especificada pela topologia de suas conexões e pelo peso e pelas fórmulas de decisão de seus nós. Uma rede pode tomar decisões mais complexas do que um neurônio sozinho. Assim, se um neurônio com duas entradas pode decidir em qual lado da linha o ponto se localiza, três neurônios podem selecionar pontos que se encontram dentro de um triângulo:



Lógica de redes neurais (*continuação*)

As redes neurais são mais eficientes e robustas se a saída é uma função que varia lentamente com os dados de entrada. Tais redes podem realizar cálculos computacionais mais gerais e são melhores para o reconhecimento de padrão. Além disso, para treinar a rede, é conveniente que a saída seja uma função diferenciável dos parâmetros. Para isso, uma função de saída de um neurônio com características abruptas é substituída por uma função degrau mais suave ou sigmóide:



função degrau função sigmóide

Uma propriedade da rede neural que lhe garante um enorme poder é que os pesos utilizados podem ser tratados como *variáveis*, e um cálculo ou processo de aprendizagem pode definir os pesos apropriados para uma tomada de decisão particular ou para um identificador de padrão. Para treinar uma rede, alimenta-se o sistema com conjuntos de amostra de entrada, cujas saídas desejadas são conhecidas, e se comparam os resultados com a resposta correta. Se os resultados observados diferem do desejado, ajustam-se os parâmetros. A topologia da rede permanece invariável durante o processo de treinamento, embora, é claro, atribuir o valor 0 a um dos pesos resulte na eliminação daquele dado de entrada.

O tipo de redes neurais que tem sido aplicado na predição de estruturas secundárias está representado na Figura 5.9.

O maior avanço na predição de estrutura secundária ocorreu com a aplicação de informações evolucionárias, o reconhecimento de que tabelas de alinhamento múltiplo de seqüências contêm muito mais informações do que seqüências individuais. A conservação da estrutura secundária entre proteínas relacionadas significa que as correlações seqüência-estrutura são muito mais robustas quando uma família, como um todo, é levada em consideração. A maioria dos métodos para a predição de estruturas secundárias, baseados em redes neurais, atualmente alimentam a camada de entrada não somente com as identidades dos aminoácidos em posições sucessivas, mas também com um perfil derivado de um alinhamento múltiplo da seqüência.

Também se provou ser útil realizar dois experimentos de redes neurais em *tandem* para utilizar as correlações observadas entre as conformações de resíduos em posições vizinhas. As predições dos estados de vários resíduos sucessivos, por uma rede semelhante à mostrada na Figura 5.9, são combinadas em uma segunda rede para a predição final.

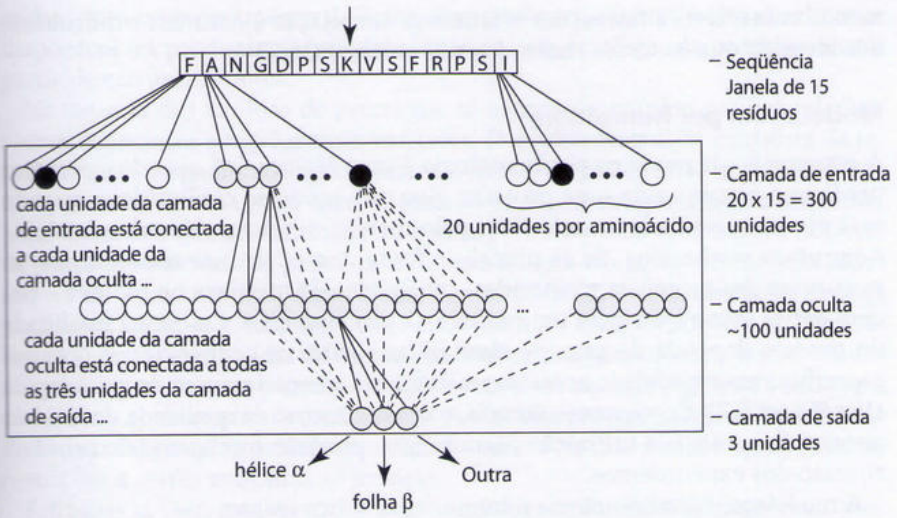


Figura 5.9 Uma rede neural aplicada à predição de estrutura secundária possui três camadas:

- A camada de entrada trabalha com uma janela de 15 resíduos ao longo da seqüência. Ou seja, ela trata uma região de 15 resíduos, prediz a estrutura secundária do resíduo central (marcado com uma seta acima do resíduo) e, então, move a janela um resíduo adiante na seqüência de aminoácidos e repete o processo. Para cada um dos 15 resíduos da janela de trabalho, há 20 nós correspondentes na camada de entrada da rede, um dos quais será disparado de acordo com o aminoácido naquela posição.
- Um camada oculta de ~100 unidades conecta a entrada com a saída. Cada nó da camada oculta é conectado a *todas* as unidades de entrada e saída; nem todas as conexões são mostradas.
- A camada de saída consiste em apenas três nós, que significam a predição do resíduo central da janela de trabalho que pode estar em uma hélice, fita, ou outra conformação.

Um teste de maturidade do método de predição é se ele pode ser totalmente automatizado. Alguns métodos computacionais produzem apenas informações aproximadas da predição de estrutura da proteína, exigindo a intervenção humana para se determinar a forma final. Outros são automáticos, e existe uma série de servidores na *web* que aceitam as seqüências e retornam as predições. O PROF, sistema que fez a predição da estrutura secundária de MutS, é um exemplo.

Uma análise permanente, completamente automática, de servidores na *web* de predição de estrutura de proteínas, incluindo mas não se limitando à estrutura secundária, é chamada EVA. EVA resulta de uma colaboração entre grupos de Nova York, San Francisco e Madri. O Protein Data Bank fornece seqüências um pouco antes de as respectivas estruturas serem liberadas, e o *software* implementado no EVA as submete a servidores de predição e avalia seus resultados. O EVA pode ser considerado como um projeto CASP contínuo, restrito a métodos que podem ser aplicados *e* avaliados automaticamente. Ver: `http://cubic.bioc.columbia.edu/eva`.

Os objetivos do EVA são monitorar os progressos na área e indicar aos usuários o melhor servidor de predição de estrutura de proteínas, em diferentes categorias. O EVA tem acesso a muito mais dados do que tem sido disponibilizado nos próprios projetos CASP. Por isso, em princípio, suas conclusões são

menos vulneráveis a flutuações estatísticas em relação à natureza e dificuldade dos alvos do que os testes realizados nos projetos CASP.

Modelagem por homologia

A construção de modelos por homologia é uma técnica útil quando queremos predizer a estrutura de uma proteína-alvo de seqüência conhecida, e quando esta proteína está relacionada com, pelo menos, outra proteína de seqüência *e* estrutura conhecidas. Se as proteínas forem estreitamente relacionadas, as estruturas das proteínas conhecidas – chamadas de matrizes ou moldes – podem servir como base para um modelo da proteína-alvo. Embora a qualidade do modelo dependa do grau de similaridade entre as seqüências, é possível especificar essa qualidade antes da realização dos experimentos de modelagem (Ver Figura 5.7). Como conseqüência, o conhecimento da qualidade do modelo necessária para a sua utilização permite uma previsão inteligente do provável sucesso dos experimentos.

A modelagem por homologia é composta de cinco etapas:

1. Alinhamento da seqüência de aminoácidos da proteína-alvo com a da proteína, ou proteínas, de estrutura conhecida. Normalmente, se observará que as inserções ou deleções localizam-se nas regiões de alças, entre elementos de estrutura secundária regular, como hélices e fitas.
2. Determinação dos segmentos da cadeia principal para representarem as regiões contendo inserções ou deleções. "Costurar" essas regiões na cadeia principal da proteína conhecida cria um modelo para a cadeia principal completa da proteína-alvo.
3. Substituição das cadeias laterais dos resíduos mutados. Para os resíduos que não mutaram, mantém-se a conformação das cadeias laterais. Os resíduos mutados tendem a manter os mesmos ângulos conformacionais das cadeias laterais, e poderiam ser modelados a partir deles. Entretanto, há métodos computacionais disponíveis que procuram entre as possíveis combinações de conformações das cadeias laterais.
4. Verificação do modelo – tanto visualmente quanto por programas – para detecção de qualquer colisão séria entre átomos. Alivie estas colisões o máximo possível, manipulando-as manualmente.
5. Refinamento do modelo por minimização de energia. A função deste passo é corrigir as relações geométricas em locais onde regiões da cadeia principal foram conectadas, e permitir que as cadeias laterais se movam um pouco a fim de se colocarem em posições geometricamente satisfatórias. Os efeitos são de fato apenas cosméticos (superficiais) – o refinamento pela minimização de energia não corrigirá erros graves no modelo.

Em grande parte, este procedimento produz "um presente", no sentido de que ele define o modelo da proteína de estrutura desconhecida fazendo alterações mínimas em uma proteína relacionada, de forma rápida e gratuita. Infelizmente, não é possível fazer melhorias substanciais. A regra básica (referindo-me, novamente, à Figura 5.7) é que se as duas seqüências tiverem, pelo menos, 40-50% de aminoácidos idênticos em um alinhamento ótimo de suas seqüências, o procedimento descrito produzirá um modelo suficientemente preciso para ser utilizado em muitas aplicações. Para proteínas muito pouco relacio-

nadas, nem o procedimento descrito nem qualquer outro método atualmente disponível irá produzir um modelo, correto nos detalhes, da proteína-alvo a partir da estrutura-molde.

Na maioria das famílias de proteínas, as estruturas contêm regiões relativamente constantes e regiões mais variáveis. O núcleo central da estrutura da família retém seu padrão de enovelamento, apesar de poder estar distorcido, mas a região mais externa pode estar enovelada de forma completamente diferente. Uma única estrutura relacionada ou molde permitirá uma modelagem razoável da porção conservada da proteína-alvo, mas falhará em gerar um modelo satisfatório da região variável. Além disso, não será fácil prever quais são as regiões variáveis e quais as constantes. Uma situação mais favorável ocorre quando várias proteínas relacionadas de estruturas conhecidas podem servir como moldes na modelagem de uma proteína-alvo. Elas revelam aquelas regiões de estrutura variável ou conservada em uma família. A distribuição observada da variabilidade estrutural entre as proteínas-molde fornece uma distribuição apropriada de restrições a serem aplicadas ao modelo.

Softwares já bem estabelecidos para modelagem por homologia encontram-se disponíveis. O SWISS-MODEL é um sítio da *web* que aceita a seqüência de uma proteína-alvo, determina se existe uma ou mais proteínas-molde para modelagem por homologia e, em caso positivo, fornece um conjunto de coordenadas para a proteína-alvo. O SWISS-MODEL foi desenvolvido por T. Schwede, M. P. Peitsch e N. Guex, agora no The Geneva Biomedical Research Institute. Outro programa de uso bem difundido, o MODELLER, foi originalmente desenvolvido por A. Šali.

Recursos na *web*: Modelagem por homologia

SWISS-MODEL (servidor de modelagem por homologia)
`http://www.expasy.ch/swissmod/SWISS-MODEL.html`

Os resultados da aplicação do SWISS-MODEL a proteínas de seqüência conhecida estão disponíveis no 3DCrunch:
`http://www.expasy.ch/swissmod/SWISS-MODEL.html`

MODELLER (*software* de modelagem por homologia)
`http://salilab.org/modeller/modeller.html`

Os resultados da aplicação do MODELLER a proteínas de seqüência conhecida estão disponíveis no MODBASE:
`http://modbase.compbio.ucsf.edu/modbase-cgi/search_form.cgi`

Para uma descrição dos sítios na *web* de genômica estrutural: Wixon, J. (2001), Structural genomics on the web, *Comp. Funct. Genomics*, **2**, 103-113.

Um exemplo de predição automática pelo SWISS-MODEL é a predição de estruturas de uma neurotoxina do escorpião-vermelho (*Buthus tamulus*) a partir da estrutura conhecida da neurotoxina relacionada do escorpião-amarelo da África do Norte (*Androctonus australis hector*). Estas duas proteínas têm 52% de resíduos idênticos no alinhamento de suas seqüências. Com tal grau de similaridade, não é de surpreender que o modelo seja muito próximo do resultado experimental, mesmo em relação às conformações das cadeias laterais (Figura 5.10).

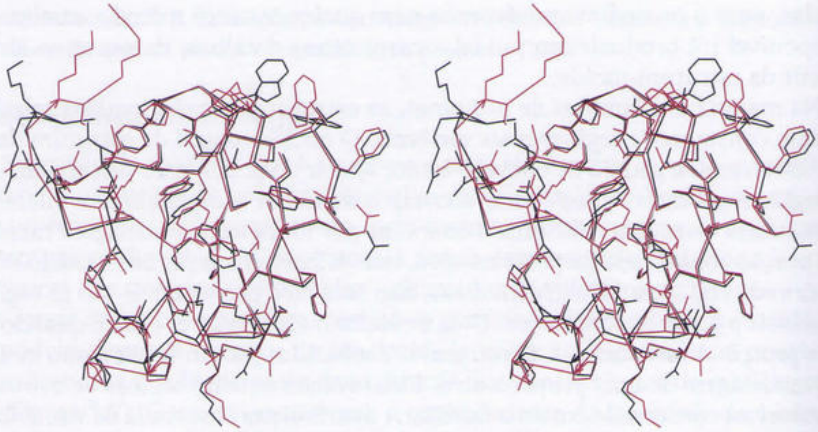


Figura 5.10 O SWISS-MODEL prediz a estrutura da neurotoxina do escorpião-vermelho [código PDB: 1DQ7] (em vermelho) a partir de uma proteína estreitamente relacionada [código PDB: 1PTX] (em preto). A predição foi feita *automaticamente*. Observe que a maioria das cadeias laterais isoladas no interior não mutou e possui conformações muito semelhantes. Algumas cadeias laterais da superfície têm conformações diferentes, e a porção C-terminal da cadeia principal está em uma posição diferente (em cima, à esquerda). Uma rede de pontes dissulfeto, que ajuda a manter a estrutura, não está representada. Entretanto, um modelo desta alta qualidade seria esperado para proteínas tão proximamente relacionadas, mesmo sem a presença destas restrições posicionais extras.

Reconhecimento de padrão de enovelamento ou topologia

A busca de uma determinada seqüência em um banco de dados de seqüências e a busca de uma determinada estrutura em um banco de dados de estruturas são problemas com soluções conhecidas. Os problemas mistos – a busca em um banco de dados de estrutura a partir de uma seqüência, ou em um banco de dados de seqüência a partir de uma estrutura – são menos diretos. Eles necessitam um método para avaliar a compatibilidade de uma dada seqüência com um dado padrão de enovelamento ou topologia.

O objetivo é extrair a essência de um conjunto de seqüências ou estruturas. Outras proteínas que seguem o mesmo padrão devem adotar estruturas similares.

Perfis 3D

Discutimos padrões e perfis derivados de alinhamentos múltiplos de seqüências e sua aplicação na detecção de homologias distantes. Uma forma de tirar vantagem das informações estruturais atualmente disponíveis, a fim de melhorar a capacidade destes métodos, é um tipo de perfil derivado de seqüências e estruturas disponíveis de uma família de proteínas.

J. U. Bowie, R. Lüthy e D. Eisenberg analisaram os *ambientes* de cada posição nas estruturas de proteínas conhecidas e os relacionaram com um conjunto de preferências dos 20 aminoácidos para estes contextos estruturais.

Dada a estrutura de uma proteína, eles classificaram o ambiente de cada aminoácido em três categorias distintas:

1. As interações tipo ligações de hidrogênio de sua cadeia principal, ou seja, sua estrutura secundária.

2. A extensão em que ele está isolado no interior ou na superfície da estrutura da proteína.
3. A natureza polar/não-polar de seu ambiente.

A estrutura secundária pode ser uma de três possibilidades: *hélice*, *folha* e *outra*. Uma cadeia lateral é considerada isolada no interior se a área de sua superfície acessível é menor do que 40 $Å^2$, parcialmente isolada no interior se sua superfície acessível está entre 40 e 114 $Å^2$, e exposta se sua superfície acessível for maior do que 114 $Å^2$. A fração de área da cadeia lateral devida a átomos polares também é medida. Os autores definem seis classes, baseadas na acessibilidade e na polaridade dos vizinhos. As cadeias laterais em cada uma das seis classes podem estar em qualquer das três classes, de acordo com sua estrutura secundária. Isso resulta em um total de 18 classes.

A atribuição de cada cadeia lateral a uma das 18 classes significa que é possível codificar uma estrutura de proteína como uma mensagem, com um alfabeto de 18 letras, chamada de **perfil da estrutura 3D**. Os algoritmos desenvolvidos para as buscas de seqüências podem, então, ser aplicados a "seqüências" de estruturas codificadas. Por exemplo, pode-se tentar alinhar duas seqüências pouco relacionadas alinhando-se os perfis de suas estruturas 3D, em vez de suas seqüências de aminoácidos. O método do perfil 3D traduz estruturas de proteínas em objetos sonda unidimensionais que não retêm explicitamente a seqüência ou a estrutura das moléculas das quais foram derivados.

Agora, como podemos relacionar o perfil de estrutura 3D ao acervo de seqüências e estruturas conhecidas? É claro que alguns aminoácidos ficarão insatisfeitos em certos tipos de posições; por exemplo, uma cadeia lateral carregada não ficaria isolada no interior de um ambiente totalmente apolar. Outras preferências não são tão claras de se apontar, e é necessário fazer uma tabela de preferências a partir de levantamentos estatísticos em uma biblioteca de estruturas bem refinadas de proteínas.

Suponha agora que nos foi dada uma seqüência e que desejamos avaliar a probabilidade de ela assumir, por exemplo, o enovelamento de globinas. A partir do perfil de estrutura 3D da estrutura conhecida da mioglobina da baleia Sperm, identificamos a classe do ambiente de cada posição na seqüência. Considere um determinado alinhamento da seqüência desconhecida com aquela da mioglobina da baleia Sperm e suponha que o resíduo na seqüência desconhecida correspondente ao primeiro resíduo da mioglobina é uma fenilalanina. A classe do ambiente no perfil de estrutura 3D do primeiro resíduo da mioglobina da baleia Sperm é: exposto, sem estrutura secundária. Pode-se avaliar a probabilidade de encontrar-se fenilalanina nessa classe de ambiente estrutural a partir da tabela de preferências de aminoácidos particulares para essa classe de perfil de estrutura 3D. (O fato de o primeiro resíduo da seqüência da mioglobina da baleia Sperm ser, na verdade, uma valina não é considerado, e na realidade esta informação não é diretamente acessível ao algoritmo. A mioglobina da baleia Sperm é representada somente pela seqüência de classes de ambiente de seus resíduos, e a tabela de preferências é obtida tomando-se a média sobre proteínas com muitos padrões de enovelamento diferentes.) A extensão deste cálculo a todas as posições e a todos os possíveis alinhamentos sem lacunas, dentro de regiões de estrutura secundária, resulta em um escore que mede quão bem a seqüência desconhecida se ajusta ao perfil da mioglobina da baleia Sperm.

Uma vantagem especial deste método é que ele pode ser automatizado, com uma nova seqüência podendo ser avaliada em relação a cada perfil 3D na biblioteca de padrões de enovelamento conhecidos, essencialmente da mesma forma que uma seqüência nova é rotineiramente comparada com um banco de dados de seqüências conhecidas.

Utilização do perfil 3D para avaliar a qualidade de estruturas

O perfil 3D derivado de uma estrutura depende, muito indiretamente, da seqüência de aminoácidos. Por isso faz sentido perguntar, não somente se é possível identificar outras seqüências de aminoácidos compatíveis com um dado enovelamento, mas se o escore de um perfil 3D para a própria seqüência referência é uma medida de compatibilidade da seqüência com a estrutura. Naturalmente, se seqüências reais não fossem compatíveis com suas próprias estruturas, poderíamos ser forçados a concluir que um método prático para examinar a relação entre seqüência e estrutura não foi obtido. Dois resultados interessantes são observados: (1) estruturas de proteínas determinadas corretamente se encaixam muito bem em seus perfis, enquanto aquelas proteínas relacionadas podem resultar em escores *mais elevados*. O perfil está extraindo as propriedades da família, e não de seqüências individuais. (2) Quando uma seqüência *não* confere com o perfil calculado a partir da estrutura experimental daquela proteína, provavelmente ocorreu um erro na determinação da estrutura. As posições no perfil que não conferem podem ser identificadas como regiões de erro.

Alinhavamento

O alinhavamento, ou *threading* em inglês, é um método de reconhecimento de enovelamento ou topologias. Dada uma biblioteca de estruturas conhecidas, e uma seqüência de uma proteína-alvo de estrutura desconhecida, essa proteína-alvo compartilha um padrão de enovelamento com alguma das estruturas conhecidas? A biblioteca de enovelamentos pode incluir apenas algumas proteínas ou todo o Protein Data Bank, ou, até mesmo, enovelamentos hipotéticos.

A idéia básica do alinhavamento é construir muitos modelos aproximados para a proteína-alvo, baseados em cada uma das estruturas conhecidas, e usar os diferentes alinhamentos possíveis das seqüências de proteínas conhecidas e desconhecidas. Esta exploração sistemática dos muitos alinhamentos possíveis dá ao alinhavamento o seu nome: imagine testar todos os alinhamentos ao se ajustar a seqüência-alvo suavemente através do delineamento tridimensional de qualquer estrutura conhecida. As lacunas devem ser permitidas nos alinhamentos, e se o ajustamento é tido como suficientemente elástico, a metáfora do alinhavamento persiste.

Tanto o alinhavamento quanto a modelagem por homologia lidam com estruturas tridimensionais inferidas por um alinhamento da seqüência-alvo com estruturas conhecidas de homólogas. A modelagem por homologia concentra-se em um conjunto de alinhamentos, e seu objetivo é um modelo bastante detalhado. O alinhavamento explora muitos alinhamentos e lida somente com modelos aproximados, muitas vezes nem mesmo construídos de forma explícita:

Modelagem por homologia	Alinhavamento
Primeiro, identifica os homólogos	Testa todos os moldes possíveis
Então, determina o alinhamento ótimo	Testa muitos alinhamentos possíveis
Otimiza um modelo	Avalia muitos modelos aproximados

Um reconhecimento de padrão de enovelamento bem-sucedido por alinhavamento exige:

1. Um método de avaliação (definição de um escore) dos modelos, de forma que possamos escolher o melhor.
2. Um método de calibração dos escores, de forma que possamos decidir se o modelo de melhor escore é verossímil.

Várias abordagens para a obtenção do escore foram testadas. Uma das mais efetivas é baseada nos padrões empíricos de resíduos vizinhos, como derivados de estruturas conhecidas. Observe a distribuição das distâncias inter-resíduos nas proteínas de estruturas conhecidas para todos os 20 × 20 pares de tipos de resíduos. Para cada par, deriva-se uma distribuição de probabilidade como função da separação no espaço e na seqüência de aminoácidos. Por exemplo, para o par Leu-Ile, considere cada resíduo Leu e Ile nas estruturas conhecidas, e para cada par Leu-Ile, registre a distância entre seus átomos Cβ e a diferença em suas posições na seqüência. A obtenção desta estatística permite estimar quão satisfatoriamente as distribuições observadas em um modelo concordam com as distribuições em estruturas conhecidas.

A equação de Boltzmann relaciona as probabilidades e energias. As aplicações comuns da equação de Boltzmann começam com uma função de energia e predizem uma distribuição de probabilidade. (Um exemplo clássico é a predição da densidade da atmosfera como função da altitude a partir da função de energia potencial gravitacional das moléculas de ar.) No alinhavamento, utiliza-se o mesmo argumento, e *deriva-se* uma função de energia *a partir* da distribuição de probabilidades. Essa função de energia é, então, utilizada para classificar os modelos de alinhavamento.

Para cada estrutura na biblioteca de enovelamentos, o processo encontra a atribuição de resíduos que produz o menor escore de energia. Embora este seja um problema de alinhamento, as interações não-locais significam que ele não pode ser resolvido por programação dinâmica.

Reconhecimento de padrão de enovelamento no CASP

Os melhores métodos de reconhecimento de padrão de enovelamento são consistentes e eficientes. Eles incluem, mas não estão limitados, a métodos baseados no alinhavamento.

As Figuras 5.11 e 5.12 mostram uma predição de A. G. Murzin, e outra predição de Bonneau, Tsai, Ruczinski e Baker, de alvos do projeto CASP 2000: ambas as proteínas de *H. influenzae* são de função desconhecida.

Cálculos de energia conformacional e dinâmica molecular

Uma proteína é uma coleção de átomos. As interações entre os átomos criam um estado único de estabilidade máxima. Encontre-o e isso é tudo!

As dificuldades computacionais nesta abordagem surgem porque (a) o modelo das interações interatômicas não é completo ou exato, e (b) mesmo que o modelo fosse exato, deveríamos enfrentar um problema de otimização de um grande número de variáveis, envolvendo não-linearidades nas funções objetivas e nas restrições, criando uma superfície de energia muito aproximada, com muitos mínimos locais. Como em um gramado de golfe com vários buracos cheios de areia, esses problemas são muito difíceis.

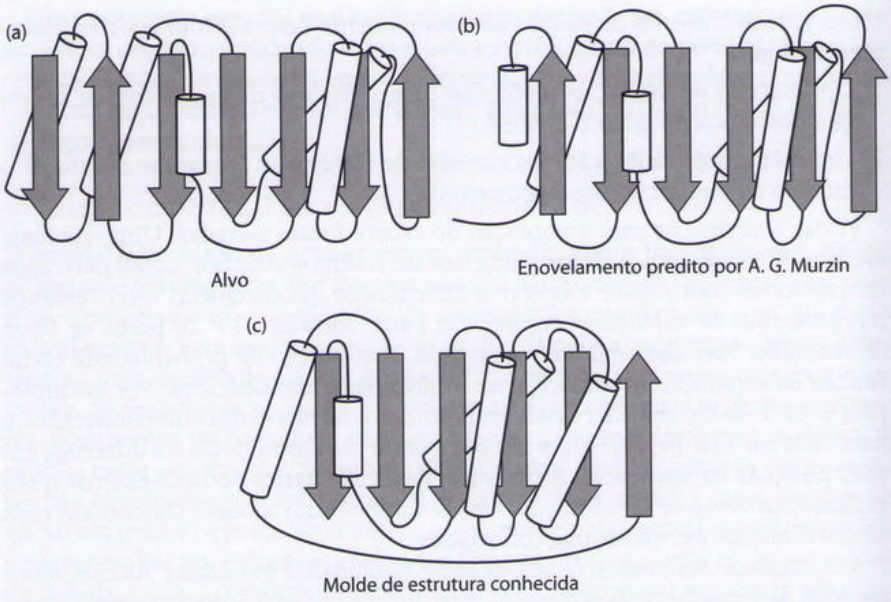


Figura 5.11 Predição de estrutura de uma proteína hipotética de *H. influenzae*. (a) Padrão de enovelamento da molécula-alvo. (b) Estrutura predita por A. G. Murzin. (c) Padrão de enovelamento da homóloga mais próxima de estrutura conhecida: uma proteína de fusão sensível à N-etilmaleimida, envolvida no transporte vesicular [Código PDB: 1NSF]. A topologia da predição de Murzin está mais próxima à do alvo do que aquela do único molde.

Figura 5.12 Predição de Bonneau, Tsai, Ruczinski e Baker de outra proteína hipotética de *H. influenzae*, baseada na glicina N-metiltransferase [código PDB: 1XVA]. Estrutura experimental em preto, estrutura predita em vermelho. Note que a maior parte da estrutura predita sobrepõe-se bem na estrutura experimental, e aquelas partes que não se sobrepõem bem possuem estruturas locais similares, mas orientação imprópria e empacotamento contra o núcleo central da proteína.

As interações entre os átomos em uma molécula podem ser divididas em:

(a) Ligações químicas primárias – interações fortes entre átomos que precisam estar bem próximos no espaço. Estas são consideradas como um conjunto fixo de interações que não são desfeitas ou formadas quando muda a conformação de uma proteína, mas, de qualquer modo, são igualmente consistentes com um grande número de conformações.

(b) Interações mais fracas que dependem da conformação da cadeia. Estas podem ser significativas em algumas conformações, mas não em outras – elas afetam conjuntos de átomos que são aproximados devido a diferentes padrões de enovelamento da cadeia.

A conformação de uma proteína pode ser especificada pela lista de átomos na estrutura, por suas coordenadas e pelo conjunto de ligações químicas entre eles (isso pode ser extraído, com uma pequena ambigüidade, a partir da seqüência de aminoácidos). Os termos utilizados na avaliação da energia conformacional tipicamente incluem:

- **Estiramento de ligação** $\Sigma_{ligações} K_r(r - r_0)^2$. Nesta expressão r_0 é a separação interatômica de equilíbrio e K_r é a constante de força do estiramento da ligação. Os valores de r_0 e K_r são dependentes do tipo de ligação química.
- **Deformação do ângulo de ligação** $\Sigma_{ângulos} K_\theta(\theta - \theta_0)^2$. Para cada átomo i que está quimicamente ligado a dois (ou mais) outros átomos j e k, o ângulo $i-j-k$ possui um valor de equilíbrio θ_0 e uma constante de força de deformação K_θ.
- Outros termos que reforçam a estereoquímica penalizam os desvios da planaridade de certos grupos, ou impõem a quiralidade correta de certos centros quirais.
- **Ângulo de torção** $\Sigma_{diedros} \frac{1}{2}V_n[1 + \cos(n\theta)]$. Para quaisquer quatro átomos conectados: i ligado a j ligado a k ligado a l, a barreira de energia de rotação por um ângulo θ do átomo l em relação ao átomo i sobre o eixo da ligação j-k é dado por um potencial periódico. V_n é a amplitude da barreira à rotação interna; n barreiras são encontradas durante uma rotação completa de 360°. Os ângulos conformacionais da cadeia principal ϕ, ψ e ω são exemplos de rotações torcíveis (ver Figura 5.2).
- **Interações de van der Waals** $\Sigma_i\Sigma_{j<i}(A_{ij}R_{ij}^{-12} - B_{ij}R_{ij}^{-6})$. Para cada par de átomos não-ligados i e j, o primeiro termo refere-se às repulsões de curta distância, e o segundo termo, às atrações de longo alcance entre eles. R_{ij} é a distância entre os átomos i e j. Os parâmetros A e B dependem do tipo de átomo.
- **Ligações de hidrogênio** $\Sigma_i\Sigma_{j<i}(C_{ij}R_{ij}^{-12} - D_{ij}R_{ij}^{-10})$. A ligação de hidrogênio é uma interação química/eletrostática fraca entre dois átomos polares. Sua magnitude depende da distância e também do ângulo de ligação. Este potencial aproximado de ligação de hidrogênio não reflete explicitamente a dependência angular da magnitude da ligação de hidrogênio; outros potenciais procuram considerar a geometria da ligação de hidrogênio de forma mais precisa.

- **Eletrostática** $\sum_i \sum_{j<i} Q_i Q_j / (\epsilon R_{ij})$. Q_i e Q_j são as cargas efetivas dos átomos i e j, R_{ij} é a distância entre eles, e ϵ é a "constante" dielétrica. Esta fórmula se aplica somente de maneira aproximada a meios que não são infinitos e isotrópicos, incluindo proteínas.
- **Solvente** Interações com o solvente água e co-solutos, como sal e açúcares, são cruciais para a termodinâmica da estrutura de proteínas. Tentativas de modelar o solvente como um meio contínuo, caracterizado, principalmente, por uma constante dielétrica, são aproximações. Com o aumento do poder computacional, hoje é possível incluir o solvente explicitamente, simulando os movimentos de uma proteína em uma caixa de moléculas de água.

Existem diversos conjuntos de potenciais de energia conformacional deste tipo ou de forma muito parecidas, e um grande esforço foi feito para ajustar os conjuntos de parâmetros. A energia de uma conformação é calculada pelo somatório destes termos sobre todos os conjuntos adequados de interações atômicas.

A função potencial satisfaz condições necessárias, mas não suficientes, para uma boa predição de estrutura. Um teste é considerar a resposta correta – uma estrutura protéica determinada experimentalmente – como a conformação inicial e minimizar sua energia a partir dali. Em geral, a maioria das funções de energia gera uma conformação minimizada que é aproximadamente 1 Å (RMSD) distante do modelo inicial. Isso pode ser entendido como uma medida da resolução de um campo de força. Outro teste seria utilizar proteínas com estruturas deliberadamente desnaturadas e minimizar suas energias conformacionais para ver se o valor de energia do mínimo local vizinho à conformação (enovelamento) correta é significativamente menor do que aquele de um mínimo local vizinho a um enovelamento incorreto. Tais testes revelam que, com base nas energias conformacionais calculadas, os múltiplos mínimos locais não podem ser diferenciados com segurança daquele que corresponde ao mínimo correto.

Os esforços para predizer a conformação de uma proteína pela minimização da energia conformacional não têm, até o momento, fornecido um método de predição da estrutura de uma proteína a partir de sua seqüência de aminoácidos. Para resolver estes problemas de ficar preso em mínimos locais de energia e da ausência de um bom modelo de interação proteína-solvente, foram desenvolvidos modelos de dinâmica molecular. A proteína, mais as moléculas explícitas de solvente, são tratadas – via campo de força – pela mecânica clássica de Newton. É verdade que isso permite a exploração de uma amostra muito maior do espaço de fase. Entretanto, como um método *a priori* (ou *ab initio*) de predição de estrutura, ele ainda não obteve um sucesso absoluto. Mas estes cálculos são computacionalmente intensivos ao extremo, e aqui, talvez mais do que em qualquer outra técnica nesta área, os avanços derivados do aumento da capacidade dos processadores terão um efeito bastante positivo.

Enquanto isso, a dinâmica molecular, quando complementada com dados experimentais, contribui metodicamente de forma muito importante para a determinação de estruturas tanto via cristalografia por difração de raios X (normalmente), quanto por ressonância magnética nuclear (sempre). Como a dinâmica molecular é integrada ao processo de determinação de estruturas? Para qualquer conformação, pode-se medir a consistência do modelo com os dados experimentais. Em cristalografia, os dados experimentais são os valores

absolutos dos coeficientes de Fourier da densidade eletrônica da molécula. Na ressonância magnética nuclear, os dados experimentais fornecem as restrições nas distâncias entre certos pares de resíduos. Contudo, tanto na cristalografia por difração de raios X (quase sempre) como na ressonância magnética nuclear, os dados experimentais subdeterminam a estrutura da proteína. Para resolver uma estrutura, deve-se procurar um conjunto de coordenadas que minimizem uma combinação dos desvios dos dados experimentais e a energia conformacional. A dinâmica molecular tem sucesso na determinação de tais conjuntos de coordenadas: a dinâmica fornece uma cobertura adequada do espaço conformacional, e a influência provida pelos dados experimentais encaminha o cálculo para a estrutura correta.

ROSETTA

ROSETTA é um programa de D. Baker e colaboradores que prediz a estrutura de proteínas a partir da seqüência de aminoácidos pela assimilação de informações sobre estruturas conhecidas. Nos projetos CASP recentes, o ROSETTA mostrou um sucesso consistente para alvos da categoria novos enovelamentos (*novel folds*). Atualmente, ele lidera a área.

ROSETTA prediz a estrutura de uma proteína por meio da geração inicial de estruturas de fragmentos, utilizando estruturas conhecidas do PDB e combinando-as. Primeiro, para cada região contígua de 3 e 9 resíduos, exemplos daquela seqüência e seqüências relacionadas são identificados em proteínas de estruturas conhecidas. Para fragmentos tão pequenos, não há nenhuma hipótese de homologia com a proteína-alvo. A distribuição de conformações dos fragmentos serve de modelo para a distribuição de possíveis conformações dos fragmentos correspondentes da estrutura-alvo.

ROSETTA explora as possíveis combinações de fragmentos utilizando simulações de Monte Carlo (ver Quadro). A função de energia possui termos que refletem a compactação, fitas β emparelhadas e o isolamento interior de resíduos hidrofóbicos em proteínas. O procedimento realiza 1.000 simulações independentes, com a estrutura inicial selecionada a partir do padrão de distribuição de conformações dos fragmentos, gerado previamente. As estruturas que resultam dessas simulações são agrupadas, e aquelas conformações localizadas no centro dos grupos maiores são apresentadas como predições da estrutura-alvo. A idéia é que uma estrutura que emerja várias vezes de muitas simulações independentes tende a possuir características favoráveis.

A Figura 5.13 mostra predições bem-sucedidas de dois alvos do projeto CASP 2000 realizadas com o ROSETTA.

ROSETTA está disponível como licença individual ou como servidor da *web*.*

LINUS

LINUS (Local Independently Nucleated Units of Structure) é um programa para a predição da estrutura de proteínas a partir da seqüência de aminoácidos, desenvolvido por G. D. Rose e R. Srinivasan. Ele é um método totalmente por primeiros princípios, sem referência explícita a qualquer estrutura conhecida ou relação seqüência-estrutura. LINUS faz o enovelamento da cadeia polipeptí-

* http://depts.washington.edu/bakerpg/

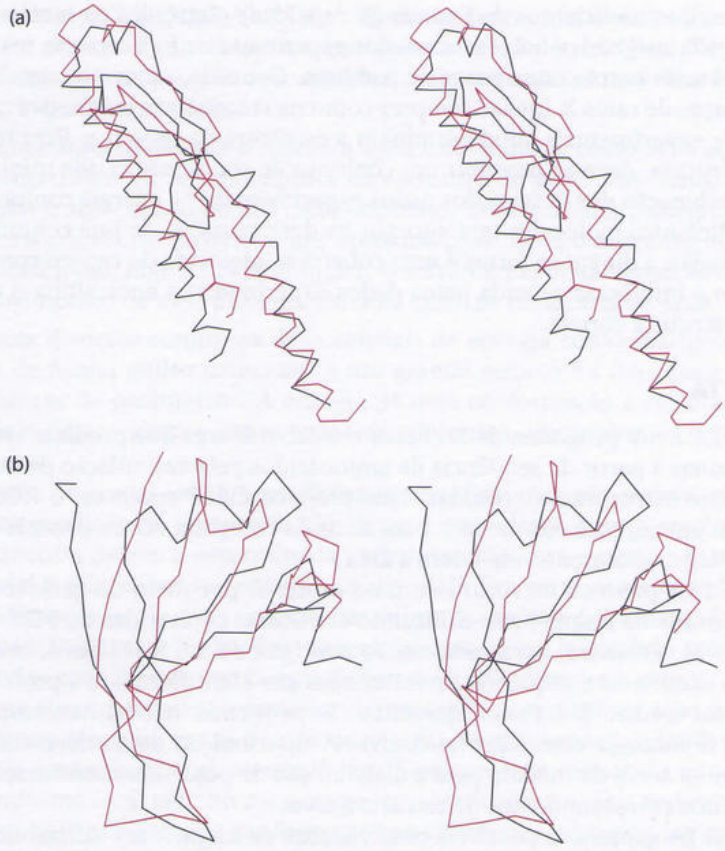


Figura 5.13 Predições pelo ROSETTA de (a) uma proteína hipotética de *H. influenzae*, (b) a metade N-terminal do domínio 1 da proteína de reparo de DNA humano XRCC4. A parte b mostra uma subestrutura selecionada contendo um segmento N-terminal de 55 resíduos (de um total de 116 resíduos). Estruturas experimentais em preto; estruturas preditas em vermelho.

dica de uma maneira hierárquica, primeiro produzindo estruturas de pequenos fragmentos, e então os reunindo em fragmentos progressivamente maiores.*

Uma característica interessante do LINUS é que a estrutura de regiões locais de uma proteína – pequenos fragmentos de resíduos consecutivos de uma seqüência – é controlada por interações locais dentro destes segmentos. Durante o enovelamento natural da proteína, cada segmento vai testar, preferencialmente, suas conformações mais favoráveis. Entretanto, essas conformações preferidas de regiões locais, até mesmo aquelas que irão, em última instância, ser adotadas no estado nativo, estão abaixo do limiar de estabilidade. As estruturas locais serão formadas transitoriamente e desfeitas muitas vezes antes que uma estrutura adjacente apropriada e interagente as estabilize. Mas no computador estamos livres para antecipar os resultados. Em uma simulação do LINUS, as estruturas favoráveis dos fragmentos locais, como determinadas pela sua freqüência de ocorrência durante a simulação,

* LINUS encontra-se disponível gratuitamente no sítio `http://roselab.jhu.edu/dist/manual/linus.html`

Algoritmos de Monte Carlo

Os algoritmos de Monte Carlo são amplamente utilizados em cálculos de estrutura de proteínas, para explorar eficientemente as conformações, e também em vários outros problemas de otimização, na procura dos mínimos de funções mais complicadas. Métodos de minimização simples, baseados no movimento "descendente" de valores de energia, falham porque suas funções acabam presas em mínimos locais, longe do estado nativo.

Em geral, os métodos de Monte Carlo utilizam números aleatórios para resolver os problemas para os quais é difícil obter uma resposta exata. O nome foi cunhado por J. von Neumann, referindo-se à aplicação de geradores de números aleatórios no famoso cassino europeu.

Para aplicar as técnicas de Monte Carlo com o objetivo de encontrar os mínimos de uma função de muitas variáveis – por exemplo, a energia mínima de uma proteína como função das variáveis que definem a sua conformação – suponha que a configuração do sistema seja especificada pelas variáveis x, e que, para quaisquer valores dessas variáveis, podemos calcular a energia conformacional, $\varepsilon(x)$. (x representa um conjunto completo de variáveis – talvez o conjunto de coordenadas atômicas de uma proteína, ou os ângulos torcíveis da cadeia principal e lateral.)

Então, o método Metropolis (inventado em 1953, supostamente em um banquete em Los Alamos, Estados Unidos) pode ser descrito em cinco passos:

1. Gera um conjunto de valores de x aleatórios para fornecer uma conformação inicial. Calcula a energia desta conformação, $\varepsilon = \varepsilon(x)$.
2. Perturba as variáveis: $x \rightarrow x'$, para gerar uma conformação próxima.
3. Calcula a energia da nova conformação, $\varepsilon(x')$.
4. Decide se *aceita* o passo: ir de $x \rightarrow x'$, ou se fica em x e tenta uma perturbação diferente:
   (a) Se a energia diminuiu, isto é, $\varepsilon = \varepsilon(x) > \varepsilon(x')$ – em outras palavras, o passo foi em uma direção a um *decréscimo* de energia (*downhill*) – sempre a aceita. A nova conformação passa a ser a conformação atual: assim, $x' \rightarrow x$ e $\varepsilon = \varepsilon(x')$.
   (b) Se a energia aumentou ou permaneceu a mesma, isto é: $\varepsilon = \varepsilon(x) \leq \varepsilon(x')$ – em outras palavras, o passo foi em direção a um *acréscimo* de energia (*uphill*) – *algumas vezes* aceita a nova conformação. Se $\Delta = \varepsilon(x') - \varepsilon(x)$, o passo é aceito com uma probabilidade $\exp[-\Delta/(kT)]$, onde k é a constante de Boltzmann, e T é uma temperatura efetiva.
5. Retorna ao passo 2.

O passo 4b é que é engenhoso. Ele tem o potencial de ultrapassar barreiras de energia, fugindo das armadilhas de mínimos locais. A temperatura efetiva, T, controla a probabilidade de que um movimento para um valor maior de energia (*uphill*) seja aceito. T não é uma temperatura física na qual queremos predizer a conformação da proteína, mas simplesmente um parâmetro numérico que controla o cálculo. Para qualquer temperatura, quanto mais positiva a diferença de energia (movimento *uphill*), menor a probabilidade de o passo ser aceito. Para qualquer valor de ε, se T é baixo,

→

→

então $\varepsilon(x)/(kT)$ será alto, e a $\exp[-\varepsilon(x)/(kT)]$ será relativamente baixa. Se T é alto, então $\varepsilon(x)/(kT)$ será baixo, e a $\exp[-\varepsilon(x)/(kT)]$ será relativamente alta. Quanto maior a temperatura, mais provável é a aceitação de um movimento em direção a uma maior energia.

Esta idéia relativamente simples provou ser extremamente eficiente, com aplicações bem-sucedidas, incluindo mas de forma alguma se limitando, aos cálculos de estrutura de proteínas.

O **recozimento simulado (*simulated annealing*)** é uma parte dos cálculos de Monte Carlo no qual a temperatura T varia – inicialmente ela é considerada alta para permitir uma exploração eficiente de conformações, e então é reduzida para levar o sistema a um estado de energia mais baixo.

transmitem suas conformações preferidas como parâmetros que influenciam as etapas subseqüentes. O procedimento aplica o efeito catraca para tornar os cálculos mais produtivos.

LINUS inicia a construção do polipeptídeo a partir da seqüência como uma cadeia estendida. A simulação prossegue com a perturbação das conformações de uma sucessão de segmentos de três resíduos escolhidos aleatoriamente e com a avaliação das energias dos resultados. Estruturas com colisões estéricas são rejeitadas imediatamente; outras contribuições energéticas são avaliadas somente em termos de interações locais. Um procedimento de Monte Carlo (ver Quadro) é utilizado para decidir se a estrutura perturbada será aceita ou se retorna à estrutura anterior. LINUS realiza um grande número destes passos. Ele amostra periodicamente as conformações dos resíduos para acumular informações estatísticas de preferências estruturais.

Os estágios subseqüentes da simulação reúnem as regiões locais em fragmentos maiores, utilizando as tendências conformacionais das regiões pequenas para guiar o processo. A janela de seqüência que controla o conjunto de interações é progressivamente aberta, a partir de pequenas regiões, até regiões maiores, para finalmente incluir a proteína inteira.

A representação do processo de enovelamento de uma proteína pelo LINUS é realística em aspectos essenciais, apesar de aproximada. Todos os átomos de uma proteína, exceto os hidrogênios, são modelados, mas a função de energia é aproximada e a dinâmica simplificada. A função de energia considera: (1) a repulsão estérica que impede a sobreposição dos átomos, (2) o agrupamento dos resíduos hidrofóbicos isolados interiores, (3) as ligações de hidrogênio, e (4) as pontes salinas.

Hoje, o LINUS é geralmente bem-sucedido na obtenção de estruturas corretas de pequenos fragmentos (tamanhos entre estrutura supersecundária e domínio) e, em alguns casos, pode reuni-los na estrutura global correta. A Figura 5.14 mostra a predição pelo LINUS do domínio C-terminal da proteína ERp29 do retículo endoplasmático de rato.

Atribuição de estruturas de proteínas a genomas

Uma seqüência genômica é uma listagem completa de uma vida em potencial. A identificação das estruturas dos produtos gênicos é o primeiro passo

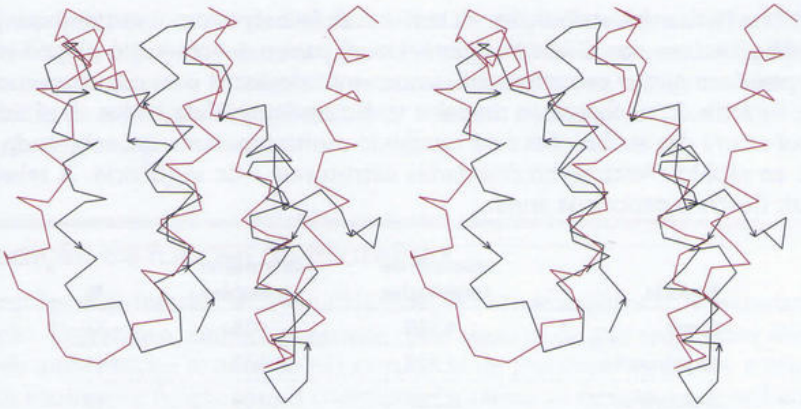


Figura 5.14 Predição pelo LINUS do domínio C-terminal da proteína ERp29 do retículo endoplasmático de rato. Estrutura experimental em preto; estrutura predita em vermelho.

no entendimento de como os organismos implementam suas informações genéticas.

Queremos compreender as estruturas das moléculas codificadas por um genoma, suas atividades individuais e interações, e a organização dessas atividades e interações no espaço e no tempo durante a vida do organismo. Queremos entender as relações entre as moléculas codificadas no genoma de um indivíduo e suas relações com aquelas de outros indivíduos e de outras espécies.

Para proteínas individuais, o conhecimento de suas estruturas é essencial para compreendermos o mecanismo de suas funções e interações. Para o organismo como um todo, o conhecimento das estruturas nos informa como o repertório dos possíveis enovelamentos de proteínas é utilizado e como ele é distribuído entre diferentes categorias funcionais em diferentes espécies. Para comparações entre espécies, estruturas de proteínas podem revelar relações invisíveis entre seqüências muito divergentes.

Vários métodos têm sido aplicados à atribuição de estruturas:

- **Determinação experimental de estruturas**. O melhor de todos!
- **Detecção de homologia entre seqüências**. Métodos sofisticados de comparação de seqüências como o PSI-BLAST ou os modelos baseados nas cadeias ocultas de Markov (Hidden Markov Models, HMM) podem identificar relações entre proteínas, tanto dentro de um mesmo organismo quanto entre espécies. Se a estrutura de alguma homóloga é conhecida experimentalmente, pelo menos o padrão de enovelamento geral da família pode ser deduzido.
- **Métodos de reconhecimento de enovelamentos ou topologias** podem atribuir padrões de enovelamento a algumas proteínas mesmo na ausência de evidência de homologia.
- Técnicas específicas detectam **proteínas de membrana** e ***coiled-coils***.*

* N. de T. *Coiled-coil* não possui nem mesmo uma tradução aproximada em português. É um tipo de motivo estrutural de proteínas composto basicamente de hélices α enroladas uma em volta da outra.

Os resultados das atribuições de estruturas fornecem um levantamento parcial de proteínas nos diferentes genomas, e, para o subconjunto de proteínas que possuem outras proteínas suficientemente similares com estrutura conhecida, as atribuições fornecem modelos tridimensionais detalhados. A extensão da cobertura das atribuições está mudando muito rapidamente, sobretudo devido ao rápido crescimento dos dados estruturais e de seqüência. A tabela a seguir ilustra o panorama atual.

Espécie	Número de seqüências	Estruturas atribuídas	%
E. coli	4.289	916	21
M. jannaschii	1.773	262	15
S. cerevisiae	6.289	1.109	18
D. melanogaster	13.687	2.990	22

(Fonte: GeneQuiz, `http://jura.ebi.ac.uk:8765/ext-genequiz/`)

O que estes resultados nos informam sobre a utilização do repertório de potenciais proteínas? Atualmente, as proteínas de estruturas conhecidas podem ser separadas em cerca de 750 tipos de enovelamentos (topologias), de um total estimado de 1.000. Uma comparação dos enovelamentos deduzidos a partir dos genomas da arqueobactéria *Methanococcus jannaschii*, da bactéria *Haemophilus influenzae* e do eucarioto *Saccharomyces cerevisiae* revelou que, de um total de 148 tipos de enovelamentos, 45 eram comuns a todas as três espécies – e, por conseqüência, provavelmente comuns à maioria das formas de vida. A arqueobactéria *M. jannaschii* possui o menor número de tipos de enovelamentos em comum (ver Figura 5.15).

Um levantamento das estruturas comuns a todas as três espécies mostrou que os cinco padrões de enovelamento mais comuns de domínios são (1) o enovelamento NTP hidrolase contendo a alça P (ou *P-loop-containing NTP*

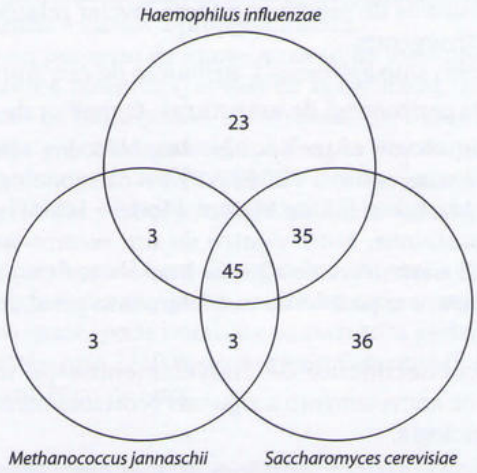


Figura 5.15 Enovelamentos compartilhados por proteínas da arqueobactéria *Methanococcus jannaschii*, da bactéria *Haemophilus influenzae* e do eucarioto *Saccharomyces cerevisiae*. (Fonte: Gerstein, M. (1997), A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure, *J. Mol. Biol.*, **274**, 562-576.)

hydrolase), (2) o domínio de ligação ao NADH, (3) o enovelamento barril TIM (*TIM-barrel*), (4) o enovelamento flavodoxina, e (5) o enovelamento de ligação à tiamina. A Gravura VI mostra a estrutura e um diagrama esquemático simplificado da topologia deste último (ver também os *Weblemas* 5.3 e 5.4). Todos são da classe α/β.

Predição de função de proteínas

A seqüência de inferência deveria idealmente fluir de seqüência → estrutura → função. Entretanto, embora possamos estar seguros de que seqüências similares de aminoácidos resultarão em estruturas de proteínas similares, a relação entre estrutura e função é mais complexa. Proteínas de estruturas semelhantes, e até mesmo de seqüências semelhantes, podem ser recrutadas para diferentes funções. Mais do que isso, assim como muitas seqüências diferentes são compatíveis com uma mesma estrutura, proteínas de enovelamentos diferentes podem ter a mesma função.

À medida que as proteínas evoluem, elas podem:

1. reter sua função e especificidade,
2. reter sua função, mas alterar sua especificidade,
3. mudar para uma função relacionada, ou uma função similar em um contexto metabólico diferente,
4. mudar para uma função completamente diferente.

As pessoas geralmente perguntam: Quanto uma seqüência ou estrutura de proteína pode modificar-se antes de alterar a sua função? A resposta é: Algumas proteínas têm múltiplas funções, de modo que elas não precisam mudar de forma alguma!

- No pato, uma lactato desidrogenase ativa e uma enolase servem como cristalinas na lente dos olhos (cristalino), embora elas não encontrem os substratos *in situ*. Por outro lado, as cristalinas são estreitamente relacionadas a enzimas, mas algumas divergências já ocorreram, com a perda da atividade catalítica. (Isso prova que a atividade enzimática não é necessária na lente dos olhos.)
- Uma proteína de *E. coli*, chamada de Do ou DegP ou HtrA, atua como uma chaperona (catalisando o enovelamento de proteínas) a baixas temperaturas, mas a 42°C ela se transforma em uma proteinase. A razão parece ser que, sob condições normais, ou de estresse térmico moderado, seu objetivo é recuperar aquelas proteínas que estão tendo dificuldade para se enovelar; sob condições mais severas de estresse térmico, quando a recuperação é impossível, seu objetivo passa a ser reciclá-las.
- Já mencionamos a enzima lipoato desidrogenase de *E. coli*, que também é uma subunidade essencial da piruvato desidrogenase, da 2-oxoglutarato desidrogenase e do complexo de clivagem de glicina.

Esses exemplos de relações estrutura-função estão no extremo de um espectro que pode incluir uma grande variedade de comportamentos.

Um problema é que não é fácil definir quantitativamente a idéia de diferença de função. Quando duas funções diferentes são mais similares uma à outra do que com duas outras funções diferentes? Em alguns casos, funções alteradas podem ocultar similaridade de mecanismos. Por exemplo, a super-

família das enolases contém várias enzimas homólogas que catalisam diferentes reações com características mecanísticas semelhantes. Este grupo inclui a própria enolase, a mandelato racemase, a enzima lactonizante de muconato I e a D-glucarato desidratase. Cada uma age extraindo um próton α de um ácido carboxílico para formar um enolato intermediário. O caminho da reação subseqüente e a natureza dos produtos variam de enzima para enzima. Estas proteínas têm estruturas globais muito similares, uma variante do enovelamento *TIM-barrel*. Diferentes resíduos no sítio ativo produzem enzimas que catalisam reações diferentes.

Divergência de função: ortólogas e parálogas

A família das serina proteinases do tipo quimotripsina inclui enzimas estreitamente relacionadas, nas quais as funções são conservadas, e uma grande variedade de homólogas divergentes que desenvolveram novas funções. A tripsina, uma enzima digestiva em mamíferos, catalisa a hidrólise de ligações peptídicas adjacentes a resíduos carregados positivamente, Arg ou Lis. (Uma **cavidade de especificidade**, uma fenda na superfície do sítio ativo, é complementar, em forma e distribuição de carga, à cadeia lateral do resíduo adjacente à ligação que será clivada.) Enzimas com seqüências, estruturas, funções e especificidades similares existem em muitas espécies, incluindo humanos, vacas, salmão do Atlântico e até mesmo *Streptomyces griseus* (Figura 5.16). A similaridade da enzima de *S. griseus* com as tripsinas de vertebrados sugere uma transferência lateral de gene. Para as enzimas dos três vertebrados, cada par de seqüências tem 64% ou mais de resíduos idênticos no alinhamento, e a homóloga de bactéria tem 30% ou mais de resíduos idênticos com as outras; todas possuem estruturas muito similares. Essas enzimas são chamadas de **ortólogas** – proteínas homólogas em espécies diferentes. (Outras homólogas de bactérias possuem seqüências bem diferentes.)

A evolução também criou proteínas relacionadas na *mesma* espécie, mas com especificidades diferentes. A quimotripsina e a elastase pancreática são outras proteínas digestivas que, como a tripsina, clivam as ligações peptídicas, mas próximo a resíduos diferentes: a quimotripsina cliva ligações adjacentes a resíduos hidrofóbicos planares volumosos (Phe, Trp) e a elastase cliva ligações adjacentes a resíduos pequenos (Ala). A mudança de especificidade é efetuada por mutações de resíduos na cavidade de especificidade. Outra homóloga, a elastase de leucócito (o objeto de pesquisa em banco de dados no Capítulo 3) é essencial para a fagocitose e a defesa contra infecção. Sob certas condições, ela é responsável pelo dano aos pulmões, levando ao enfisema.

Proteínas homólogas na mesma espécie são chamadas de **parálogas**. A tripsina, a quimotripsina e a elastase pancreática atuam na digestão dos alimentos. Outro conjunto de parálogas medeia a cascata de coagulação do sangue. Apesar de todas serem proteinases, os requisitos para ativação e controle são bem distintos para a digestão e para a coagulação sangüínea, e as famílias divergiram e se tornaram especializadas nestas respectivas funções.

Algumas homólogas da tripsina desenvolveram funções completamente novas:

- A haptoglobina perdeu a sua atividade proteolítica. Ela age como uma chaperona, evitando agregações não desejadas de proteínas. A haptoglobina forma um complexo compacto com fragmentos de hemoglobina liberados de eritrócitos, com vários efeitos benéficos, como a prevenção da perda de ferro.

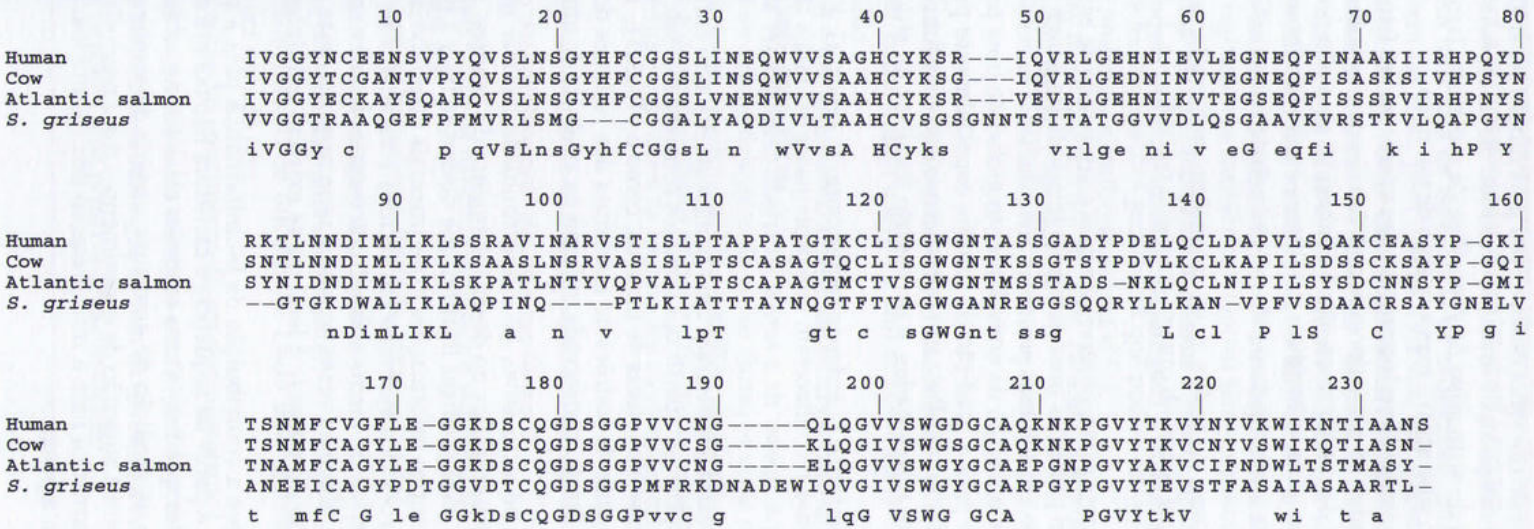


Figura 5.16 Alinhamento múltiplo de seqüências de tripsinas de humanos, vaca, salmão do Atlântico e *Streptomyces griseus*. Nas linhas abaixo do bloco de alinhamento, as letras maiúsculas indicam os resíduos absolutamente conservados, e as letras minúsculas indicam aqueles resíduos conservados em três das quatro seqüências (na maioria das vezes, mas nem sempre, a seqüência de *S. griseus* é a exceção).

- A serina proteinase de rinovírus desenvolveu uma função separada, independente, de formar o complexo de iniciação da síntese de RNA, usando resíduos do lado oposto ao do sítio ativo de proteólise da molécula. Isso não é uma modificação de um sítio ativo – mas a criação de um novo.
- Subunidades homólogas às serina proteinases aparecem em fatores de crescimento relacionados ao plasminogênio. A função dessas subunidades na atividade do fator de crescimento não é conhecida, mas não pode ser uma função proteolítica, pois os resíduos catalíticos essenciais foram perdidos.
- Uma glicoproteína anticongelante do peixe antártico é homóloga ao tripsinogênio.
- A proteína "imune" de inseto, a *scolexin*, é uma homóloga distante das serina proteinases que induzem a coagulação de hemolinfa em resposta à infecção.

Na família da quimotripsina, observamos uma conservação de estrutura com funções similares em proteínas proximamente relacionadas e uma divergência progressiva de função em algumas proteínas distantemente relacionadas.

A conclusão é que o padrão de enovelamento global de uma proteína não é um guia confiável para a predição da função, especialmente para aquelas homólogas muito distantes. Para uma predição correta da função em proteínas distantemente relacionadas, é necessário focalizar no sítio ativo. Por exemplo:

- J. F. Bazan e R. Fletterick e, independentemente, P. Argos, G. Kamer, M. J. Nicklin e E. Wimmer reconheceram que proteinases 3C virais são homólogas à quimotripsina, apesar de a serina da tríade catalítica ter sido trocada por cisteína.
- W. R. Taylor e L. Pearl identificaram uma homologia distante entre proteinases retrovirais e aspárticas a partir dos resíduos conservados Asp, Tre e Gli.

Da mesma forma que bibliotecas de motivos, como o PROSITE, tais abordagens partem diretamente de padrões de assinatura dos resíduos do sítio ativo na seqüência para a função conservada, mesmo na ausência de uma estrutura experimental.

Concentrando-se no sítio ativo, há a oportunidade de usar métodos semelhantes àqueles empregados no desenvolvimento de ligantes, no planejamento de fármacos, para predizer ligantes que devem se ligar a proteínas. É importante utilizar outras informações experimentais disponíveis, tais como distribuição de padrões de expressão em tecidos e catálogos de interações de proteínas. Tentativas de medir a função diretamente, por exemplo, pelo "nocaute de genes", algumas vezes levarão a uma resposta, mas são improdutivas se o fenótipo do nocaute é letal ou se existem múltiplas proteínas com aquela função.

Parece provável que a contribuição da bioinformática para a predição da função de proteínas a partir da seqüência e estrutura não poderá ser feita por um algoritmo simples que leve a uma resposta clara (como se acredita que, um dia, será o caso da predição de estrutura a partir da seqüência). Metas mais plausíveis seriam a sugestão de experimentos mais produtivos e a contribuição da bioinformática para a interpretação dos resultados. Estas, sim, são metas que valem a pena.

Descoberta e desenvolvimento de fármacos

Perguntar em uma sala de aula, cheia de estudantes, quantos deles estariam vivos hoje sem ter feito, pelo menos uma vez, um tratamento com remédios para curar uma doença séria é uma experiência, no mínimo, interessante. (Isso, ignorando aquelas doenças evitadas pela vacinação.) Ou, ainda, perguntar aos estudantes quantos de seus avós ainda vivos estariam tendo uma qualidade de vida reduzida sem um tratamento com fármacos. As respostas são impressionantes. Eles manifestam preocupações com novas cepas de microrganismos infecciosos resistentes a antibióticos. Outros fármacos têm como alvo proteínas humanas, para lidar com disfunções de proteínas ou para ajustar os controles reguladores. É necessário o desenvolvimento de novos fármacos que, em combinação com as informações genômicas que podem aumentar sua especificidade, deverão prolongar e melhorar nossas vidas.

Entretanto, não é fácil ser um fármaco. Para um composto químico ser qualificado como um fármaco, ele deve ser:

1. seguro
2. eficaz
3. estável – tanto química quanto metabolicamente
4. entregável, isto é, atingir o alvo – o fármaco deve ser absorvido e chegar a seu sítio de ação
5. disponível – por isolamento a partir de recursos naturais ou por síntese
6. novo, ou seja, patenteável

As etapas no desenvolvimento de novos fármacos estão resumidas no Quadro adiante. O processo envolve pesquisa científica, ensaio clínico para provar eficácia e segurança, e os aspectos legais e econômicos muito importantes, incluindo proteção de patente e estimativa de retorno dos altos investimentos necessários.

Etapas no desenvolvimento de um novo fármaco

1. Entendimento da natureza biológica e sintomas de uma doença. Ela é causada por:
   - um agente infeccioso – bactéria, vírus ou outro?
   - um veneno de origem não-biológica?
   - uma proteína mutante do paciente?
2. Desenvolvimento de um ensaio. Dado um candidato a fármaco, ele pode ser testado por:
   - seu efeito no crescimento de um microrganismo?
   - seu efeito em células crescidas em cultura de tecidos?
   - seu efeito em animais que sofrem da doença ou em um análogo?
   - sua ligação a uma proteína-alvo conhecida?

→

Etapas no desenvolvimento de um novo fármaco (*continuação*)

3. Existe um agente eficaz de alguma fonte natural conhecida folclórica ou tradicionalmente? Se sim, vá para a etapa 6.
4. Identificação de um alvo molecular específico, normalmente uma proteína. Determinação de sua estrutura experimentalmente ou por modelagem.
5. Obtenção de uma idéia geral de que tipo de molécula se encaixaria no sítio da molécula-alvo. Existe um substrato ou inibidor conhecido?
6. Identificação de um *composto líder*: um composto químico que apresenta a atividade biológica desejada em *qualquer* nível de medição. Um composto líder é um obstáculo a ser vencido; encontrar tais compostos e subseqüentemente modificá-los são atividades bastante distintas.
7. Desenvolvimento do composto líder: estudo extenso de variantes do composto, com o objetivo de gerar todas as propriedades desejadas e aumentar a atividade biológica.
8. Testes pré-clínicos, *in vitro* e com animais, para provar eficácia e segurança. Neste ponto, o fármaco pode ser patenteado. (A princípio, tenta-se adiar o processo de patente o máximo possível por causa do período finito da patente, pois ainda faltam várias etapas demoradas antes de o fármaco poder ser comercializado.)
9. Nos Estados Unidos: a submissão de um formulário para análise de novos fármacos, o Investigational New Drug Application, ao FDA (Food and Drug Administration). Isso é seguido de três fases de ensaios clínicos.
10. Fase I dos ensaios clínicos. Testes de segurança do composto em voluntários sadios. Determinação de como o corpo reage ao fármaco – como ele é absorvido, distribuído, metabolizado e excretado. Os resultados sugerem um intervalo de dosagem segura.
11. Fase II dos ensaios clínicos. Testes de eficácia do composto contra uma doença em aproximadamente 200 pacientes voluntários. Ele cura a doença ou alivia seus sintomas? Calibração da dosagem.
12. Fase III dos ensaios clínicos. Teste, em aproximadamente 2.000 pacientes, para demonstrar, conclusivamente, que o composto é melhor do que o melhor protocolo de tratamento conhecido. Estes são ensaios clínicos aleatórios duplo-cegos, tanto contra um placebo como contra um fármaco utilizado no momento. Estes ensaios são extremamente caros; não é incomum abortar um projeto antes de se chegar a esta etapa, caso os ensaios clínicos da fase II identifiquem efeitos colaterais ou eficácia insatisfatória.
13. Registro do fármaco (New Drug Application) no FDA, com os dados que garantem sua segurança e eficácia. A aprovação do FDA permite a comercialização do fármaco. Somente neste momento é que o fármaco começa a gerar um retorno financeiro.
14. Os estudos da fase IV, após a aprovação do FDA e comercialização, envolvem o monitoramento continuado dos efeitos do fármaco que podem ser avaliados como uma experiência mais abrangente do seu uso. Novos efeitos colaterais podem surgir em algumas classes de pacientes, levando a restrições no uso do fármaco, ou possivelmente ao seu recolhimento.

Para desenvolver um fármaco, inicialmente precisamos escolher uma doença de interesse. Você pode querer estudar o que é conhecido sobre suas possíveis causas, seus sintomas, sua genética, sua epidemiologia, sua relação com outras doenças – humanas e animais – e todos os tratamentos conhecidos. Assumindo que o potencial de utilização do fármaco justifica o tempo despendido, o custo e os esforços necessários para o seu desenvolvimento, você está pronto para começar.

A partir da doença de interesse, você deve escolher uma proteína-alvo. Aproximadamente 500 proteínas são alvos de fármacos conhecidos.

Você deve desenvolver um ensaio adequado que detecte sucesso já na fase inicial. Se o alvo é uma proteína conhecida, sua ligação ao fármaco pode ser medida diretamente. O efeito de um fármaco potencialmente antibacteriano pode ser testado verificando o seu efeito no crescimento do patógeno. Os efeitos de alguns compostos poderiam ser testados no crescimento de células eucarióticas em cultura de tecidos. Se um animal de laboratório é suscetível à doença, os compostos podem ser testados nestes animais. Contudo, os compostos podem apresentar diferentes efeitos em animais e humanos. Por exemplo, o tamoxifeno, uma droga muito utilizada atualmente no tratamento do câncer de mama, foi desenvolvido originalmente como pílula anticoncepcional. De fato, o tamoxifeno é um bom contraceptivo para ratos, mas *promove* a ovulação em mulheres.

O composto líder

Um objetivo nos estágios iniciais do desenvolvimento de fármacos é a identificação de um ou mais compostos líderes. Um composto líder é qualquer substância que apresenta a atividade biológica que se procura. Isso demonstra que existe um composto com, no mínimo, algumas das propriedades desejadas.

Existem diversas formas de se encontrar compostos líderes:

1. Serendipidade ou descoberta por acaso – a penicilina é o exemplo clássico.
2. Inspeção de fontes naturais. "Procure e ache" é o que move os químicos medicinais. Algumas vezes, os remédios tradicionais apontam para uma fonte de compostos ativos. Por exemplo, a digitalina foi isolada das folhas da dedaleira, que tinha sido utilizada como tônico do coração. (Por que não continuar simplesmente com o uso do remédio tradicional? O isolamento do princípio ativo torna possível regular a dosagem e explorar as variantes.) Aproximadamente metade dos fármacos utilizados hoje é baseada em produtos naturais.
3. Estudo do que se sabe sobre substratos, inibidores e o mecanismo de ação, para então selecionar compostos potencialmente ativos a partir destas propriedades.
4. Consideração de fármacos eficazes contra doenças similares.
5. Triagem em larga escala. Técnicas de química combinatorial permitem o teste em paralelo de um grande número de compostos relacionados. Uma técnica especialmente utilizada com polipeptídeos é a exibição por fagos (*phage display*).
6. Ocasionalmente, a partir de efeitos colaterais de fármacos existentes. O minoxidil (2,4-diamino-6-piperidino-pirimidina-3-óxido), originalmente desenvolvido para ser um anti-hipertensivo, mostrou ser um indutor de crescimento de cabelo. O Viagra, originalmente desenvolvido como um remédio para o coração, é outro exemplo.

7. Triagem experimental. O US National Cancer Institute, nos Estados Unidos, testou dezenas de milhares de compostos. (A triagem de variantes é também muito importante *depois* que um composto líder é encontrado.)
8. Triagem computacional e planejamento *ab initio* computacional.

A descoberta de compostos líderes promoveu outros tipos de atividades de pesquisa. Muitos variantes de compostos líderes devem ser testados para melhorar sua eficácia e para incluir novas propriedades essenciais. Por exemplo, um composto que se liga ao seu alvo *in vitro* não é um bom fármaco a menos que ele possa chegar ao alvo *in vivo*. O encontro de um fármaco com seu alvo dentro do corpo necessita da capacidade de ele ser absorvido e transportado. Isto exige estabilidade no "armazenamento" e estabilidade metabólica. Ele deve possuir um perfil de solubilidade adequado – um fármaco deve ser suficientemente solúvel em água para ser absorvido, mas não tão solúvel a ponto de ser excretado imediatamente; ele deve ser, na maioria dos casos, suficientemente lipossolúvel para atravessar membranas, mas não tão lipossolúvel a ponto de ser simplesmente armazenado nos depósitos de gordura do corpo. Deve também existir uma rota sintética adequada para produzir o composto em quantidade.

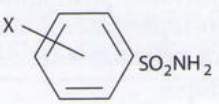
Melhorando o composto líder: Relações Quantitativas Estrutura-Atividade (QSAR)

Para qualquer composto com atividade farmacológica, compostos similares normalmente possuem atividades parecidas, mas variam em sua potência e especificidade. Baseados em um composto líder, os químicos investigam um grande número de moléculas parecidas de forma a otimizar as propriedades farmacológicas desejadas. Para uma busca sistemática, seria muito importante o entendimento de como as variações nas características estruturais e físico-químicas da família de moléculas estão relacionadas com suas propriedades farmacológicas. O problema é que existem muitos descritores diferentes para caracterizar moléculas. Eles incluem características estruturais, como natureza e distribuição dos substituintes; características experimentais, como solubilidade em solventes aquosos e orgânicos, ou momentos de dipolo; e características calculadas computacionalmente, como cargas parciais dos átomos.

As relações quantitativas estrutura-atividade (QSAR – *Quantitative Structure-Activity Relationships*) constituem métodos para predizer a atividade farmacológica de um conjunto de compostos, a partir de relações entre as características moleculares e a atividade farmacológica, baseados em dados experimentais. Este método foi desenvolvido por C. Hansch e colegas na década de 1960 e tem sido amplamente utilizado.

C. Hansch, J. McClarin, T. Klein e R. Langridge utilizaram os métodos QSAR para estudar inibidores da anidrase carbônica. A anidrase carbônica é uma enzima que catalisa a reação $CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^-$. As aplicações clínicas dos inibidores da anidrase carbônica incluem diuréticos, tratamento de pressão alta intra-ocular no glaucoma, pela supressão da secreção de humor aquoso (líquido interno dos olhos), e agentes antiepilépticos. As pessoas que fazem escaladas em grandes altitudes utilizam os inibidores da anidrase carbônica para aliviar os sintomas do mal-estar agudo causados pela altitude.

Medidas da ligação da anidrase carbônica a 29 fenilsulfonamidas do tipo



(X representa um conjunto de substituintes do anel, variáveis em estrutura e em posição) demonstraram que a constante de ligação estava relacionada com a constante σ de Hammett do substituinte eletrônico, um descritor que representa uma medida da força receptora ou doadora de elétrons do substituinte; com o coeficiente de partição octanol-água, *P*, da forma não-ionizada do ligante; e da localização da substituição (*orto* ou *meta*):

$$\log K \approx 1{,}55\sigma + 0{,}65 \log P - 2{,}07I_1 + 3{,}28I_2 + 6{,}94$$

onde K = constante de ligação, $I_1 = 1$ se X estiver na posição *meta* ou 0, caso contrário, e $I_2 = 1$ se X estiver na posição *orto* ou 0, caso contrário. Os substituintes X eram do tipo –alquil, ou –COO-alquil, ou –CONH-alquil.

Este tipo de correlação tem duas implicações:

1. Um grande número de compostos pode ser avaliado no computador e aqueles que forem julgados como melhores, testados experimentalmente.
2. É possível visualizar o sítio de ligação a partir da análise dos seguintes parâmetros:
   - O coeficiente de σ positivo, implicando que os substituintes receptores de elétrons são preferidos, sugere que a forma ionizada da porção $-SO_2NH_2$ se liga ao íon Zn no sítio ativo da anidrase carbônica.
   - O coeficiente de log *P* positivo sugere uma interação hidrofóbica entre a proteína e o ligante.
   - Os coeficientes negativos de I_1 e I_2 sugerem impedimentos estéricos com substituintes nas posições *orto* ou *meta*.

As estruturas da anidrase carbônica ligada a inibidores confirmam estas conclusões (ver *Weblema* 5.9).

A bioinformática na descoberta e no desenvolvimento de fármacos

A computação e a recuperação rápida de informações contribuem em várias etapas dos projetos de descoberta e desenvolvimento de fármacos. Isso inclui identificação da molécula-alvo, planejamento, análise e melhoramento de ligantes, além de seleção e triagem *in silico* de bibliotecas de ligantes. Sistemas de informação também são importantes na organização dos dados teóricos de predição, dos planejamentos de experimentos e análise dos dados. D. Searls chama esse equilíbrio íntimo entre teoria e experimentação de "ciclos seco e molhado" (*Wet-dry cycles*).*

Seleção do alvo

Para desenvolver um fármaco contra uma determinada doença, é necessário selecionar uma proteína associada à doença, de forma que ele seja terapeuticamente interessante para afetar sua função ou expressão. Novos métodos de geração de dados em grande escala, especialmente de seqüências de genomas e de padrões de expressão de proteínas, fornecem uma fonte valiosa de material para a identificação de potenciais alvos de fármacos. **Genômica e proteômica diferenciais**, as comparações entre humanos ou animais sadios e doentes, podem indicar que proteína em particular está faltando, não está funcional, está mal regulada ou expressa somente nas células afetadas. As informações sobre os complexos proteína-proteína tornam possível investir não somente em uma proteína, mas em uma interação proteína-proteína específica.

* N. de T. Molhado se refere aos experimentos *in vitro* (*wet*) ou *in vivo* realizados em bancadas de laboratórios. Seco (*dry*) se refere aos experimentos *in silico*.

O conhecimento de genomas virais e procarióticos auxilia na identificação de alvos para fármacos contra doenças infecciosas. Rotas metabólicas específicas de microrganismos, e as proteínas que delas participam, são de bastante interesse. Uma droga que interaja com estes alvos tem menos chance de interagir com uma homóloga humana, o que pode levar a uma redução ou eliminação dos efeitos colaterais. Proteínas com seqüências similares em diferentes espécies de bactérias oferecem a possibilidade de se desenvolver antibióticos de amplo espectro. Ao contrário, a duplicação de genes alerta para uma potencial redundância de funções, com conseqüente dessensibilização à inibição pelo fármaco. O conhecimento da velocidade de evolução relativa de diferentes proteínas, incluindo taxas de transferência horizontal de genes, indica a estabilidade esperada de uma terapia contra o desenvolvimento de cepas resistentes.

O comprometimento de uma grande indústria farmacêutica com um alvo envolve um grande investimento de recursos. Os lucros esperados para um fármaco bem-sucedido exercem uma grande influência na escolha dos alvos pesquisados. A análise da história dos fármacos que atualmente rendem grandes lucros sugere que a previsão dos retornos econômicos não é uma ciência exata. Atualmente, mesmo os esforços da bioinformática, com o apoio de investimentos bem generosos, são mais baratos do que o trabalho de laboratório. A possibilidade de que cálculos poderão melhorar as predições e aumentar os lucros está por trás da adesão à bioinformática pelas indústrias farmacêuticas, além de contribuições puramente científicas da bioinformática ao desenvolvimento de fármacos. Esta contribuição à previsão econômica é especialmente importante quando uma companhia considera projetos de alto risco, como aqueles que pretendem desenvolver fármacos contra uma nova classe de alvos. Tais projetos devem competir com atividades de mais baixo risco, como a tentativa de melhoramento de um fármaco de sucesso de uma empresa competidora.

Predição de um composto líder

Os métodos de predição de ligantes, apropriados como compostos líderes para o desenvolvimento de fármacos, podem ser divididos em indutivos e dedutivos.

Os métodos indutivos dependem das correlações entre as afinidades conhecidas de algum conjunto-teste de compostos e as propriedades moleculares que caracterizam uma biblioteca de possíveis ligantes. Essas características envolvem propriedades estruturais, como tamanho, geometria, distribuição de carga e grupos funcionais específicos, incluindo doadores e receptores de ligação de hidrogênio. Elas incluem qualidades gerais "do tipo fármaco" (*drug-like*), como solubilidade em solventes aquosos e orgânicos, uma via de administração simples, uma distribuição adequada nos tecidos do corpo e uma taxa de conversão metabólica apropriada. Os químicos medicinais aplicam um teste equivalente ao teste do pato: se ele anda como um fármaco, nada como um fármaco e "grasna" como um fármaco, então é bem provável que ele seja um fármaco. As características relevantes de compostos são compiladas em um **vetor de atributos** utilizado para comparar a similaridade geral entre compostos de afinidade conhecida e uma biblioteca completa de compostos. A necessidade de organização, codificação, armazenamento e busca de informações sobre pequenas moléculas criou uma nova área de pesquisa, a **quimioinformática**, que complementa a bioinformática nas aplicações ao desenvolvimento de fármacos.

Os métodos dedutivos são aplicáveis se o sítio de ligação na proteína-alvo é conhecido ou pode ser inferido. Entretanto, como a afinidade de ligação e a especificidade representam somente dois dos requisitos para um composto líder

– e reconhecidamente essenciais – é necessário combinar os métodos dedutivos com a correlação com as propriedades desejáveis, assim como na abordagem puramente indutiva. Os ensaios de ligação em sistemas purificados fornecem apenas uma pequena idéia sobre o comportamento de um composto, como um fármaco, em seu contexto biológico. A bioinformática tem uma contribuição a fazer na integração das informações disponíveis da biologia celular e molecular, parasitologia e farmacologia, ajudando a fazer uma ponte entre os experimentos *in vitro* e as atividades terapêuticas.

Modelagem molecular no desenvolvimento de fármacos

Um problema central no desenvolvimento de fármacos é a identificação de um composto que se ligará forte e especificamente à proteína-alvo. Uma ligação forte é necessária para a eficácia a baixas concentrações. A especificidade é necessária para minimizar os efeitos colaterais.

Se a estrutura do alvo é conhecida experimentalmente, é possível utilizar a modelagem molecular diretamente no planejamento do ligante. Se a estrutura do alvo é desconhecida, um modelo do sítio de ligação deve ser criado a partir de evidências indiretas, e o planejamento de ligantes é, conseqüentemente, mais difícil. Desenvolver ligantes sem a estrutura da molécula-alvo é como tentar pegar um assaltante de banco a partir da descrição de uma testemunha ocular. Desenvolver ligantes para um alvo de estrutura conhecida é como tentar pegar esse assaltante de banco com uma imagem clara do vídeo de uma câmera de segurança.

Os objetivos da modelagem molecular aplicada ao desenvolvimento de fármacos incluem:

- Idealmente: sugestão de um composto líder que já demonstre afinidade e especificidade razoáveis. Isso é um feito raro.
- Análise de compostos que reconhecidamente se ligam à molécula-alvo. O entendimento das interações importantes serve como guia para planejar e testar potenciais ligantes e para selecionar características estruturais para serem incluídas na síntese combinatória de bibliotecas de compostos. No caso de projetos antibacterianos e antivirais, um modelo do complexo proteína-ligante pode dar algumas idéias de quão fácil seria para o patógeno desenvolver resistência por meio de mutações, o que diminui a afinidade.
- Identificação do farmacóforo é a identificação de subestruturas comuns a vários compostos que compartilham uma atividade farmacológica, ou pelo menos se ligam ao mesmo sítio em uma proteína. A hipótese é que existe um conjunto de átomos comuns dentro das estruturas que é responsável pela ligação. O problema computacional de identificar farmacóforos a partir de um conjunto de compostos é semelhante ao do alinhamento estrutural de um conjunto de proteínas homólogas. No entanto, embora ligantes sejam muito menores do que proteínas, os problemas combinatórios são mais severos, pois não temos a ordem linear dos resíduos na proteína (ver página 255). A dedução das propriedades farmacofóricas está integrada com os métodos QSAR para filtrar bibliotecas de compostos na busca de candidatos a ligantes.
- Triagem *in silico*: a predição de afinidade, mesmo qualitativamente, pode sugerir candidatos a ligantes a partir de bibliotecas de estruturas químicas. Os resultados podem ser utilizados tanto para determinar as prioridades para testes experimentais quanto para abordagens mais amplas de triagem computacional de bibliotecas, com base nas características correlacionadas com propriedades químicas e farmacológicas favoráveis. Vários leitores deverão tomar conheci-

Ver Quadro Docagem molecular: predição da geometria e afinidade do ligante

mento da utilização em todo o mundo de telas protetoras de computadores (*screensavers*) para pesquisar por possíveis fármacos.* Mais de 3,5 milhões de computadores, em mais de 200 países, participaram deste projeto. Eles contribuíram com um total de mais de 450.000 anos de capacidade de CPU.

- **Melhoramento do composto líder**: uma vez identificado um composto que se liga a uma proteína-alvo, ainda que com baixa afinidade e especificidade, a modelagem interativa pode sugerir modificações que melhorem o seu ajuste ao alvo. A síntese e o teste daqueles compostos que, por predição, devem apresentar uma melhor afinidade, e até mesmo a resolução das estruturas cristalinas de seus complexos, podem direcionar a busca por melhores compostos. A modelagem é normalmente acoplada à química combinatória e à triagem experimental de bibliotecas de compostos.

Docagem molecular: predição da geometria e afinidade do ligante

A docagem molecular é a predição da ligação do ligante. Ela inclui a predição da ligação de pequenas moléculas a proteínas e a ligação proteína-proteína. Os objetivos da docagem molecular são (1) identificar o sítio de ligação na proteína e determinar a posição e orientação do ligante, e (2) estimar a afinidade da ligação.

(1) Identificação do modo de ligação A docagem de pequenas moléculas a proteínas requer o ajuste do ligante em uma região de uma proteína de estrutura conhecida. A região ou sítio de ligação deve ser previamente conhecido, ou deve ser necessário testar muitos modos de ligação diferentes entre ligante e proteína para predizer o sítio de ligação ótimo.

A base da docagem molecular é a identificação da complementaridade em tamanho, forma, distribuição de carga, polaridade e potencial para interações hidrofóbicas e ligações de hidrogênio. Um complicador é a possibilidade de flexibilidade em ambas as partes. Moléculas orgânicas pequenas, contendo muitas ligações simples, têm um alto grau de flexibilidade conformacional. (Projetistas de fármacos adoram estruturas moleculares com anéis e pontes.) Muitas proteínas apresentam mudanças conformacionais mediante associação com o ligante. Portanto, a estrutura experimental de uma proteína não-ligada não pode ser considerada como um alvo flexível para a docagem. Contudo, a consideração da flexibilidade complica significativamente os cálculos de docagem molecular.

Moléculas de água nas interfaces representam outra dificuldade. Elas podem contribuir para a complementaridade da superfície e prover pontes entre a proteína-alvo e o ligante através de ligações de hidrogênio.

(2) Estimativa de afinidade É difícil estimar as afinidades absolutas. Entretanto, a docagem molecular comparativa pode fornecer informações úteis sobre as afinidades *relativas*. Uma função de escore adequada, que pode predizer o ranque de diferentes ligantes em relação a uma ordem aproximada de afinidade, permite a seletividade e a determinação de prioridades nos testes experimentais. Tais esquemas de escore podem ser *ab initio* – baseados nos tipos de campos de força descritos nas páginas 277-278 – ou empíricos. De forma semelhante, a docagem comparativa de um ligante a várias proteínas pode predizer a especificidade da interação.

→

* http://www.chem.ox.ac.uk/curecancer.html

Compare:

Cálculos de docagem molecular	Informação obtida
1 ligante – 1 proteína	modo de ligação, estimativa de afinidade
muitos ligantes – 1 proteína	ranqueamento das afinidades de uma série de potenciais ligantes
1 ligante – muitas proteínas	predição de especificidade

A docagem e o ranqueamento (escore) são etapas importantes na filtragem de candidatos de uma biblioteca numerosa de possíveis ligantes e o teste de bancada. Um afunilamento típico deste processo ocorre como mostrado a seguir:

tamanho total da biblioteca	10^{12} compostos
depois de alguns filtros	10^{5}
docagem molecular	10^{4}
ranqueamento (escore)	10^{3}
visual	10-100 para teste experimental

Dois estudos de caso ilustram o conjunto de técnicas químicas e biológico-moleculares envolvidas no desenvolvimento de fármacos e mostram algumas similaridades e alguns contrastes interessantes. Eles envolvem duas famílias de analgésicos bem conhecidos – vulgarmente chamados de *"pain-killers"* – exemplificados pela morfina e pela aspirina. Os dois grupos de compostos têm diferentes mecanismos de ação, diferentes potências e diferentes espectros de efeitos colaterais.

Estudo de Caso 5.1: Desenvolvimento de analgésicos baseados na morfina*

A morfina e a codeína são alcalóides naturais contidos no látex do ópio papoula (*Papaver somniferum*) (Figura 5.17). Os efeitos farmacológicos são conhecidos desde a antigüidade. A química moderna explorou e desenvolveu muitas variantes. A heroína foi sintetizada em 1874 (Figura 5.17). Mais hidrofóbica do que os compostos naturais, a heroína atravessa a barreira sangue-cérebro mais facilmente, proporcionando um início de ação mais rápido.

Tanto a codeína quanto a heroína são metabolizadas, gerando morfina, a forma ativa. A codeína é, portanto, um exemplo de um **pró-fármaco**: um agente inativo que é transformado em um agente ativo. A conversão depende de um citocromo, o CYP2D6, que está ausente em 5-10% dos brancos e em 1-3% dos afro-americanos e asiáticos, para os quais a codeína é ineficaz.

* Coop, A. & MacKerell, A. D. Jr. (2000), The future of opioid analgesics, *Amer. J. Pharm. Educ.*, **66**, 153-156.

Estudo de Caso 5.1 (*continuação*)

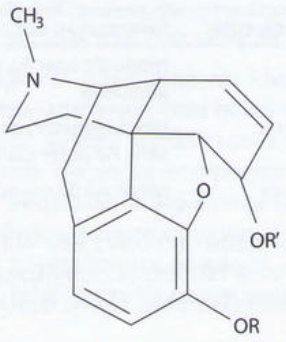


Figura 5.17 A morfina, a codeína e a heroína possuem estruturas que diferem somente nos substituintes em duas posições:

Composto	R	R´
Morfina	–H	–H
Codeína	$-CH_3$	–H
Heroína	$-COCH_3$	$-COCH_3$

A morfina e a codeína têm sido aplicadas na medicina e em cirurgias como analgésicos, fármacos para aliviar as dores intensas. Os efeitos colaterais incluem passividade e euforia, bem como dependência química e vício. Por isso, os desenvolvedores de fármacos há tempos sonham com um composto que alivie a dor sem os efeitos colaterais nocivos. É claro que não havia nenhuma garantia de que isso seria possível.

Variantes sintéticas da morfina permitem a correlação entre efeitos biológicos e estrutura química.

Uma alternativa é tentar simplificar a estrutura. Os objetivos são (1) deduzir o farmacóforo mínimo necessário para atividade, e (2), se possível, separar as partes da estrutura responsáveis pelo alívio da dor daquelas responsáveis pela dependência. A morfina, a codeína e a heroína são compostos rígidos que contêm cinco anéis fusionados. O levorfanol difere da morfina pela perda da ponte de oxigênio (eliminação do anel tetraidrofurano) e um dos grupos hidroxila (Figura 5.18). Ele é um analgésico mais potente do que a morfina, mas ainda causa dependência. O benzomorfano, a ciclazocina e a pentazocina não possuem o anel cicloexeno (Figura 5.19). Os efeitos de dependência destes compostos são menores dos que os da morfina e do levorfanol. O demerol, que possui o anel cicloexeno aberto, e a metadona, que *não* possui anéis fusionados, mantêm a atividade analgésica e compartilham ainda menos subestruturas em comum com a morfina.

A partir destas estruturas, pode-se deduzir o farmacóforo mostrado na Figura 5.20.

Em contraste com a simplificação da molécula para identificação do farmacóforo, tentativas de aumentar a especificidade preservaram o farmacóforo, mas tornaram a molécula mais complexa. Alguns sucessos foram al-

cançados. A etorfina e a buprenorfina, descobertas na década de 1960, são analgésicos muito mais potentes do que a morfina (a etorfina é usada na sedação de animais de grande porte) e possuem baixo potencial de dependência (ver Figura 5.21). Na verdade, o uso clínico mais importante da buprenorfina é no tratamento de dependência de drogas, e não na analgesia.

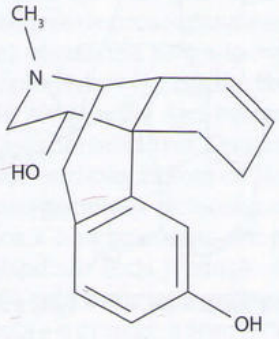


Figura 5.18 Estrutura do levorfanol.

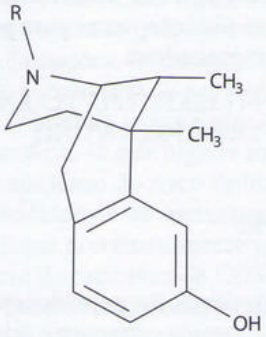


Figura 5.19 Estruturas do benzomorfano: R = CH_3; ciclazocina: R = CH_2-cp (cp = ciclopropano); pentazocina: R = $CH_2CH{=}C(CH_3)_2$.

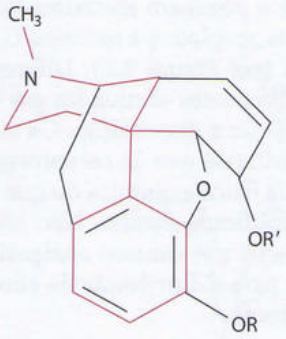


Figura 5.20 Farmacóforo (em vermelho) derivado da comparação estrutural entre os derivados de morfina. (De acordo com A. D. MacKerell, Jr.)

Estudo de Caso 5.1 (*continuação*)

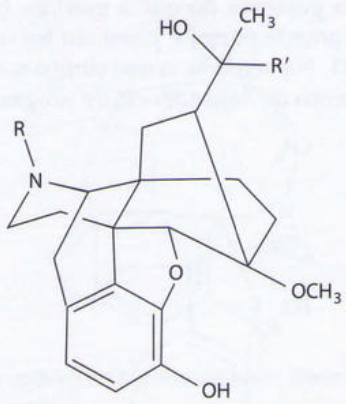


Figura 5.21 Estruturas da etorfina: R = CH_3 e R′ = C_3H_7; e da buprenorfina: R = CH_2-cp e R′ = *t*-butila.

Esta investigação de variantes foi realizada antes da identificação dos receptores naturais. Agora, sabemos que os alvos naturais de ação da morfina e de moléculas relacionadas são receptores para peptídeos endógenos chamados endorfinas. Eles compreendem:

β-endorfina	YGGFMTSEKSQTPLVTLFKNAIIKNAYKKGE
Dinorfina	YGGFLRRIRPKLKWDNQ

e seus produtos de clivagem:

Met-encefalina	YGGFM
Leu-encefalina	YGGFL

A morfina é, portanto, um composto **peptideomimético** natural, um não-peptídeo que compartilha a mesma estrutura e atividade de um peptídeo.

Várias classes de receptores são conhecidas, incluindo os tipos μ, κ e δ, e um quarto tipo, recentemente descoberto, chamado ORL-1 (ORL = *Opiate-Receptor Like*, ou receptor semelhante ao receptor de ópio). Suas seqüências de aminoácidos possuem aproximadamente 50-70% de identidade. Eles são receptores acoplados à proteína G, semelhantes em estrutura à bacteriorodopsina (ver Figura 4.7). Diferentes ligantes – naturais e sintéticos – possuem diferentes afinidades por diferentes receptores e diferentes cinéticas de ligação e dissociação. Os alvos naturais da morfina são os receptores μ. Acredita-se que os receptores μ tendam a estar mais envolvidos na dependência física e química do que os receptores κ, embora esta afirmação esteja simplificada demais. Não obstante, isso sugere que, para a produção de fármacos que causem analgesia com efeitos colaterais reduzidos, deve-se olhar para a *distribuição* de afinidade do composto nos diferentes tipos de receptores.

Estudo de Caso 5.2: Planejamento de fármacos assistido por computador: inibidores específicos da prostaglandina ciclooxigenase 2

As prostaglandinas são uma família de compostos naturais que medeiam uma grande variedade de processos fisiológicos. As aplicações farmacológicas incluem a própria utilização das prostaglandinas e, ao contrário, de drogas que bloqueiam a síntese de prostaglandinas. A prostaglandina E_2 (dinoprostona) é utilizada na obstetrícia para a indução do parto. A aspirina, o ibuprofeno, o acetaminofeno (tylenol) e outros fármacos **antiinflamatórios não-esteróides** (NSAIDs, *Nonsteroidal Anti-Inflammatory Drugs*) são efetivos contra artrite e doenças relacionadas (ver Quadro na página 303). Eles alcançam este efeito inibindo enzimas da rota de síntese das prostaglandinas, especificamente, das prostaglandinas ciclooxigenases. Um efeito colateral bem conhecido da aspirina é o sangramento das paredes do estômago. Isso ocorre porque as prostaglandinas (cuja produção é inibida pela aspirina) impedem a secreção de ácido pelo estômago, promovendo a formação de uma camada de muco que recobre e protege o revestimento do estômago.

A aspirina e outros NSAIDs inibem *duas* prostaglandinas ciclooxigenases muito relacionadas, chamadas de COX-1 e COX-2. (Infelizmente, a mesma abreviação é utilizada para as citocromo oxidases 1 e 2.) A COX-1 é expressa constitutivamente no revestimento do estômago. A COX-2 é induzida e regulada em resposta a inflamações. Isso sugere que uma droga que inibisse a COX-2, mas não a COX-1, conservaria a atividade desejada dos NSAIDs, mas reduziria os efeitos colaterais indesejados. [Nota adicionada no momento da impressão: descobriu-se que alguns inibidores de COX-2 podem estar relacionados com o aumento do risco de doenças cardiovasculares.]

As seqüências de aminoácidos e as estruturas cristalinas das COX-1 e COX-2 são conhecidas. (Estas proteínas apresentam 65% de identidade). A Figura 5.22 apresenta parte da estrutura da COX-1 acetilada por um análogo da aspirina, o ácido 2-bromoacetoxibenzóico (aspirina bromada no grupo metila da porção acetila). A porção salicilato se liga próximo. O efeito é o bloqueio da entrada do sítio ativo. A maioria dos NSAIDs se liga covalentemente à enzima, mas não a modifica.

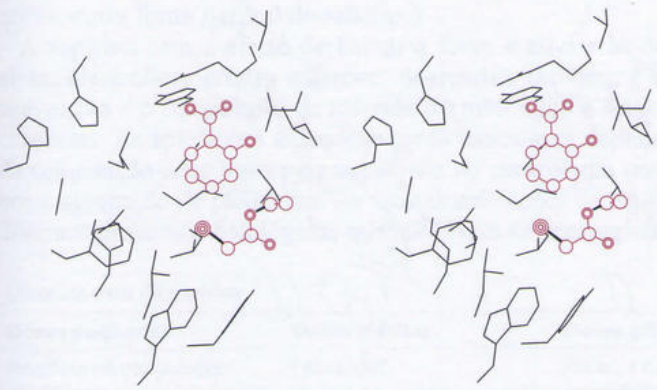


Figura 5.22 O sítio de ligação de um análogo da aspirina, o ácido 2-bromoacetoxibenzóico, na COX-1. O ligante reagiu com a proteína, transferindo seu grupo bromoacetil para a cadeia lateral da serina 530. A proteína é representada em forma de palitos, em preto. O análogo da aspirina é representado como esfera e palitos, em vermelho.

→

Estudo de Caso 5.2 (*continuação*)

A Figura 5.23 apresenta a mesma figura com a região correspondente em COX-2 sobreposta. Você consegue enxergar regiões de diferenças estruturais que poderiam representar uma informação adicional para o planejamento de fármacos seletivos? A Figura 5.24 apresenta a região da COX-2 onde o inibidor seletivo SC-558 (1-fenilsulfonamida-3-trifluormetil-5-parabromofenilpirazol, sintetizado pela Searle). A partir da Figura 5.25, podemos observar por que o SC-558 não pode inibir a COX-1. Existiria um impedimento estérico com a cadeia lateral da isoleucina, que corresponde à valina na COX-2.

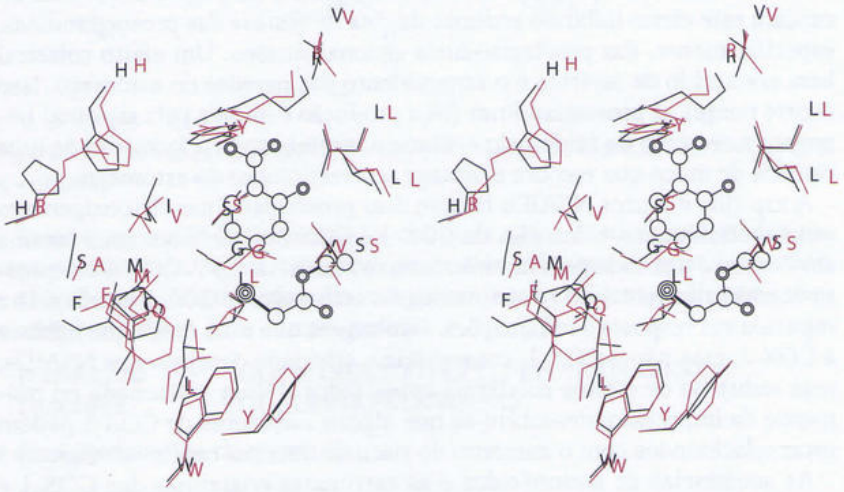


Figura 5.23 O sítio de ligação de um análogo da aspirina, o ácido 2-bromoacetoxibenzóico, em preto, na COX-1, e os resíduos homólogos da COX-2, em vermelho. Você consegue ver quais os espaços não ocupados que existem no sítio que poderiam acomodar um ligante maior? Você consegue ver qualquer diferença na seqüência que poderia ser explorada para o desenho de inibidores que se ligariam à COX-2 (em vermelho), mas *não* à COX-1 (em preto)?

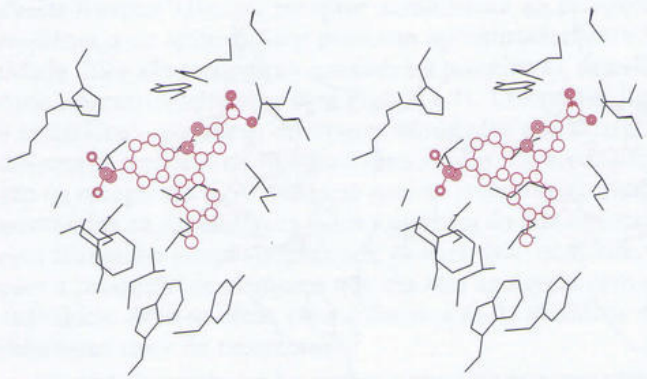


Figura 5.24 O sítio de ligação na COX-2 (em preto) para um inibidor *seletivo* de COX-2, o SC-558 (1-fenilsulfonamida-3-trifluormetil-5-parabromofenilpirazol) (em vermelho).

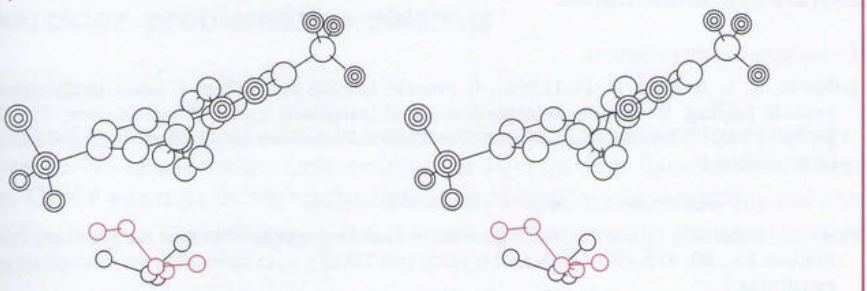


Figura 5.25 O inibidor SC-558 e o resíduo de COX-1 (em preto, isoleucina) e COX-2 (em vermelho, valina) que parece ser responsável pela seletividade. O SC-558 não pode se ligar à COX-1 porque existem contatos estéricos desfavoráveis entre ele e a isoleucina.

Aspirina

A aspirina é um dos remédios populares mais antigos e, dos científicos, o primeiro. Há pelo menos 5.000 anos, Hipócrates notou a eficácia dos preparados de folhas e casca de salgueiro para aliviar a dor e baixar a febre. O ingrediente ativo, a salicina, foi purificado em 1828, e sintetizado em 1859, por Kolbe. O mecanismo de sua ação não era conhecido e, de fato, continuou desconhecido até que J. Vane e colegas, nos anos de 1970, descobriram que a aspirina atua bloqueando a síntese de prostaglandinas. O desconhecimento do mecanismo de ação nunca foi um impedimento para a sua utilização.

Há um século, o salicilato de sódio foi utilizado no tratamento de artrite. Como a irritação estomacal era um efeito colateral sério, F. Hoffman procurou reduzir a acidez do composto pela síntese do ácido acetilsalicílico, ou aspirina. A aspirina foi a primeira droga sintética, a qual iniciou a indústria farmacêutica moderna. (O nome salicina vem do nome em latim para salgueiro, *salix*, e o nome da aspirina vem de "a" de acetil, e "spir" da planta *spirea*, outra fonte natural de salicina.)

A aspirina tem o efeito de baixar a febre e aliviar as dores. Em doses altas, ela é eficaz contra a artrite. A aspirina também é utilizada para a prevenção e o tratamento de infartos do miocárdio e acidentes vasculares cerebrais. As aplicações a doenças cardiovasculares dependem da inibição da coagulação sangüínea pela supressão do controle da prostaglandina sobre a agregação de plaquetas. As muitas aplicações da aspirina refletem os diversos processos fisiológicos que envolvem as prostaglandinas.

Diversos usos da aspirina

Doses pequenas	Doses médias	Doses grandes
Interfere na coagulação sangüínea	Febre/dor	Reduz a dor e a inflamação causadas pela artrite e doenças relacionadas

Leituras recomendadas

Enovelamento de proteínas

Baldwin, R. L. & Rose, G. D. (1999), Is protein folding hierarchic? I. Local structure and peptide folding. II. Folding intermediates and transition states, *Trends Biochem. Sci.*, **24**, 26-32; 77-83. [Uma introdução ao pensamento atual sobre estabilidade e enovelamento de proteínas.]

Alinhamento de estrutura e relações estrutura-atividade

Holm, L. & Sander, C. (1995), Dali: a network tool for protein structure comparison, *Trends Biochem Sci.*, **20**, 478-480. [Descreve o programa DALI e suas aplicações ao alinhamento de estruturas.]

Smith, T. F. (1999), The art of matchmaking: sequence alignment methods and their structural implications, *Structure Fold. Des.*, **7**, R7-R12. [Descrição do trabalho de análise de seqüência e estrutura.]

Relações entre seqüências, estruturas e funções

Das, R., Junker, J., Greenbaum, D. & Gerstein, M. B. (2001), Global perspectives on proteins: comparing genomes in terms of folds, pathways and beyond, *The Pharmacogenomics Journal*, **1**, 115-125.

Galperin, M. Y. & Koonin, E. V., Sequence-Evolution-Function/Computational Approaches in Comparative Genomics (Boston: Kluwer 2003).

O estado da arte em modelagem por homologia e sua aplicação na genômica estrutural

Tramontano, A. (2004), Integral and differential form of the protein folding problem, *Physics of Life Revs.*, **1**, 103-127.

Guex, N., Diemand, A. & Peitsch, M. C. (1999), Protein modelling for all, *Trends. Biochem. Sci.*, **24**, 364-367.

Peitsch, M. C., Schwede, T. & Guex, N. (2000), Automated protein modelling–the proteome in 3D, *Pharmacogenomics*, **1**, 257-266. [O que falta para resolver o problema da genômica estrutural.]

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. (2003), SWISS-MODEL: An automated protein homology-modeling server, *Nucl. Acids Res.*, **31**, 3381-3385. [Descrição do SWISS-MODEL.]

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., Šali, A. (2000), Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291-325.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M. S., Davis, F. P., Stuart, A. C., Mirkovic, N., Rossi, A., Martí-Renom, M. A., Fiser, A., Webb, B., Greenblatt, D., Huang, C. C., Ferrin, T. E. & Šali, A. (2004), MODBASE, a database of annotated comparative protein structure models, and associated resources, *Nucl. Acids Res.*, **32**, D217-D222.

Outros métodos de predição de estrutura de proteínas

Bonneau, R. & Baker, D. (2001), Ab initio protein structure prediction: Progress and Prospects, *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173-189. [Revisão recente dos métodos de predição de estrutura, pelos autores do método de maior sucesso.]

Clássicos que ainda valem a leitura

Kauzmann, W. (1959), Some factors in the interpretation of protein denaturation, *Adv. Protein Chem.*, **14**, 1-63.

Richards, F. M. (1977), Areas, volumes, packing and protein structure, *Annu. Rev. Biophys. Bioeng.*, **6**, 151-176.

Chothia, C. (1984), Principles that determine the structure of proteins, *Annu. Rev. Biochem.*, **53**, 537-572.

Richards, F. M. (1991), The protein folding problem, *Scientific Amer.*, **264(1)**, 54-57, 60-63.

Exercícios, problemas e *weblemas*

Exercícios

5.1 O calor de sublimação do gelo é igual a 51 kJ mol^{-1} à temperatura de congelamento. No estado sólido, cada molécula de H_2O faz duas ligações de hidrogênio. Qual é a energia de uma única ligação de hidrogênio água-água?

5.2 Quais pares são ortólogas, quais são parálogas e quais não são nem um nem outro?

(a) Hemoglobina α humana e hemoglobina β humana.

(b) Hemoglobina α humana e hemoglobina α de cavalo.

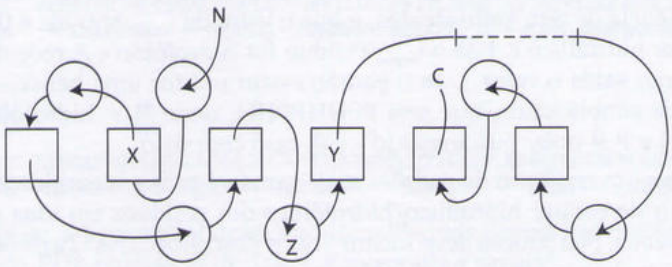
(c) Hemoglobina α humana e hemoglobina β de cavalo.

(d) Hemoglobina α humana e hemoglobina ζ humana.

(e) As proteinases quimotripsina humana e trombina humana.

(e) As proteinases quimotripsina humana e actinidina do quiuí.

5.3 Em uma fotocópia da Gravura VI, indique as localizações na estrutura que correspondem aos pontos X, Y e Z do seguinte diagrama.



5.4 Em uma fotocópia da Figura 5.8b, realce a região da hélice 3_{10} que não foi predita como hélice.

5.5 Qual das seguintes representações indica a topologia correta – ordem correta das fitas na seqüência e orientação no espaço 3D – da folha β da Figura 5.8b?

(a) ↑↑↑↑ 1 2 3 4 (b) ↑↓↑↓ 3 4 2 1 (c) ↑↑↓↑ 1 3 2 4

5.6 Na predição da estrutura da proteína hipotética de *H. influenzae*, Figura 5.11: (a) Quais são as diferenças no padrão de enovelamento entre a proteína-alvo e a estrutura experimental da proteína relacionada ou molde? (b) Quais são as diferenças no padrão de enovelamento entre a estrutura predita por A. G. Murzin e o alvo? (c) Quais são as diferenças no padrão de enovelamento entre a estrutura predita por A. G. Murzin e aquela determinada experimentalmente? Em quais aspectos a predição de Murzin constitui uma melhor representação do padrão de enovelamento em comparação com a estrutura experimental?

5.7 Desenhe a estrutura química da aspirina e do ácido 2-bromoacetoxibenzóico.

5.8 Muitas proteínas de patógenos possuem homólogas em humanos. Suponha que você possua um método que compare os determinantes de especificidade nos sítios de ligação de duas proteínas homólogas. Como você poderia utilizar esse método com o objetivo de selecionar alvos adequados para o desenvolvimento de fármacos?

5.9 Na rede neural ilustrada na página 266 (figura inferior), quantos parâmetros – pesos variáveis e limites – podem ser ajustados, assumindo um procedimento de decisão linear?

5.10 Qual é a interpretação geométrica de um neurônio que aceita duas entradas x e y e "dispara" se, e somente se, $x + 2y \geq 2$?

5.11 Esboce um neurônio com duas entradas x e y, sendo que cada uma delas pode ter qualquer valor numérico, que emitirá uma saída igual a 1 se, e somente se, o valor da primeira entrada for maior ou igual ao valor da segunda. Qual é a interpretação geométrica desse neurônio?

Problemas

5.1 Na tabela de seqüências alinhadas de domínios ESTs (ver Problema 1.1): (a) Quais são os membros mais similares e mais distantes da família? (b) Suponha que a estrutura experimental seja conhecida somente para a primeira seqüência. Para quais outras seqüências você seria capaz de construir um modelo com um RMSD ≤ 1,0 Å para 90% ou mais dos resíduos?

5.2 Desenhe uma rede neural que aceite oito entradas, cada uma com o valor 0 ou 1, com a interpretação de que as oito entradas correspondem a resíduos de uma seqüência de oito aminoácidos, e que o valor da $i_{\text{ésima}}$ entrada é 0 se o $i_{\text{ésimo}}$ resíduo for hidrofílico e 1 se o $i_{\text{ésimo}}$ resíduo for hidrofóbico. A rede deve apresentar como saída o valor 1, se o padrão estrutural for uma hélice – por uma questão de simplicidade, que seja PPHHPPHH, onde H = hidrofóbico (não-carregado) e P = polar ou carregado – e 0, caso contrário.

5.3 Escreva um conjunto de padrões mais razoável para a identificação de hélices a partir do caráter hidrofílico/hidrofóbico dos resíduos em uma seqüência de 10 resíduos. Seu padrão deve incluir "tipos desconhecidos" (*wild cards*) – posições que podem ser tanto hidrofóbicas quanto hidrofílicas, ou correlações entre diferentes posições. Generalize o problema anterior desenhando uma rede neural para detectar esses padrões mais complexos.

5.4 Nós e os computadores podemos fazer lógica com aritmética. Defina: 1 = VERDADEIRO e 0 = FALSO. Esboce neurônios simulados com duas entradas, cada qual podendo ter somente os valores 0 ou 1, e um processo de decisão linear para o disparo, no qual (a) a saída é a função lógica AND (E) das duas entradas, e (b) a saída é a função lógica OR (OU) das duas entradas. (c) Qual é a rede neural mais simples, com cada neurônio possuindo um processo de decisão linear para o seu disparo, que produz como sua saída o EXCLUSIVO OR das duas entradas (o EXCLUSIVO OR é VERDADEIRO se pelo menos uma das entradas é VERDADEIRO, e FALSO se nenhuma ou ambas as entradas forem VERDADEIRO). Isso pode ser feito com uma única camada? Se não, qual é o número mínimo de camadas necessárias para a rede?

5.5 Modifique o programa PERL para desenhar uma roda helicoidal (páginas 252-253) de forma que aminoácidos diferentes sejam representados com a mesma fonte, mas apareçam em cores diferentes, como GAST em ciano; CVILFYPMW em verde; HNQ em magenta; DE em vermelho; e KR em azul.

5.6 Análise de agrupamento hidrofóbico. Suponha que uma região de uma proteína forme uma hélice α. Para representar a sua superfície, imagine-se torcendo a seqüência em uma hélice α (mesmo que de fato ela forme uma fita de

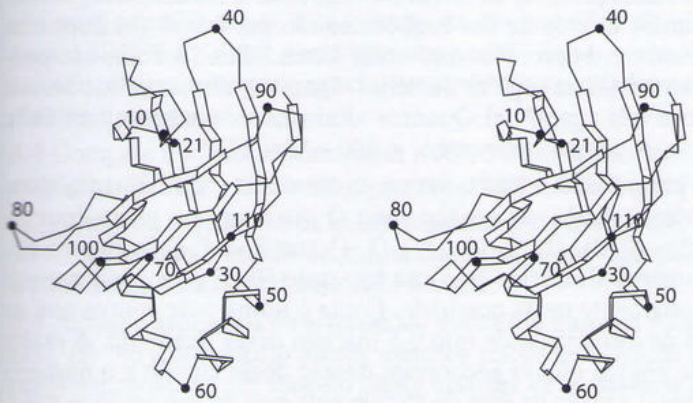
folha ou alça na estrutura nativa). Então, "borre com uma tinta ou cor" a superfície desta hélice e role-a em uma folha de papel para colocar os nomes dos resíduos. Rolando a hélice duas vezes, todas as faces podem ser visíveis simultaneamente. A partir deste diagrama, áreas hidrofóbicas nas superfícies das hélices podem ser identificadas. Dessa forma, é possível tentar predizer quais regiões da seqüência realmente formam hélices na estrutura nativa. Comparações com grupos hidrofóbicos também podem ser utilizadas para detectar relações mais distantes.

Escreva um programa PERL para produzir este tipo de diagrama.

5.7 No CASP 2000 (CASP4), um dos alvos na categoria para a qual não se conhecia nenhum padrão de enovelamento similar era o domínio N-terminal da proteína humana de ligação de DNA XRCC4 (resíduos 1-116). A estrutura secundária predita por B. Rost, utilizando o método PROF (*Profile Based Neural Network Prediction*), foi a seguinte (um H sob um resíduo significa que aquele resíduo foi predito como estando em uma hélice, um E significa que aquele resíduo foi predito como estando em uma conformação estendida, ou fita, e "–" significa outro):

```
                1         2         3         4         5         6
                0         0         0         0         0         0
Sequence    MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
Prediction  ---EEEEEEE-----HHHHHH-HHHHHHH--EEEEEE------EE---HHHHHHHHHHHH

                                          1         1
                7         8         9         0         1
                0         0         0         0         0
Sequence    MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKV
Prediction  HHH-HHHHHHHHHHHH-----EEEEEE-----EEEEE------EEEE-----HHHH
```

A estrutura experimental deste domínio, liberada depois das predições, e depositada no PDB [código PDB: 1FU1], é mostrada a seguir:



A atribuição de estruturas secundárias a partir do arquivo PDB é:

Estrutura secundária	Intervalo dos resíduos
Hélice	27-29, 49-59, 62-75
Folha 1	2-8, 18-24, 31-37, 42-48, 114-115
Folha 2	84-88, 95-101, 104-111

(a) Calcule o valor de Q3, a porcentagem de resíduos atribuídos corretamente como hélice (H), fita (E) e outro (–).

(b) Em uma fotocópia da figura da XRCC4, identifique, com cores diferentes, as regiões *preditas* como hélices e fitas.

(c) A partir do resultado de (b), quantas regiões preditas como hélices se sobrepõem com as hélices da estrutura experimental? Quantas regiões preditas como fitas se sobrepõem com as fitas da estrutura experimental?

5.8 No CASP4, o grupo de Bonneau, Tsai, Ruczinski e Baker fizeram a predição da estrutura tridimensional completa da proteína XRCC4 (resíduos 1-116). A predição da estrutura secundária derivada do modelo deles é a seguinte (H = hélice, E = fita (estendida), e "–" = outro):

```
                 1         2         3         4         5         6
                 0         0         0         0         0         0
Sequence     MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
Prediction   ----E--EEEE---EEEE--EHHHHHHHH----EEEE--EEEE-----HHHHHHHHHHHH

                 7         8         9         0         1
                 0         0         0         0         0
Sequence     MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKV
Prediction   HHH---HHHHHHHHHHH-----EEEEEEE--EEEEEEE------HHHH----HHHH
```

(a) Qual é o valor de Q3 para esta predição? (b) Neste caso, qual método fornece o melhor resultado, de acordo com o parâmetro Q3, para a predição da estrutura secundária: a rede neural que produz somente a predição da estrutura secundária, ou a predição completa da estrutura tridimensional?

5.9 Escreva e teste programas em PERL que implementem a rede neural mostrada na página 267.

5.10 Suponha que você esteja tentando avaliar, utilizando a abordagem de alinhavamento, se uma seqüência de tamanho M deve apresentar o mesmo padrão de enovelamento de uma proteína de estrutura conhecida e tamanho $N > M$. (a) Quantos alinhamentos diferentes das seqüências são possíveis? (b) Suponha que metade dos resíduos da proteína conhecida forma hélices α e não são permitidas lacunas dentro destas regiões de hélice. Quantos alinhamentos dessas seqüências são possíveis agora? (c) Quantos alinhamentos existem, em cada um dos casos acima, se assumirmos que $N = 200$ e $M = 150$?

5.11 Escreva um programa em PERL para calcular valores aproximados de π pelo método de Monte Carlo, de acordo com: O quadrado no plano com os vértices nas posições (0, 0), (1, 0), (0, 1), e (1, 1) tem área 1. Calcule uma série de pares de números aleatórios (x, y) no intervalo [0, 1] para gerar pontos distribuídos aleatoriamente neste quadrado. Conte o número de pontos que se encontram dentro de um círculo de raio 0,5 inscrito nesse quadrado. A razão entre o número de pontos que se encontram dentro desse círculo e o número total de pontos é igual à razão da área do círculo pela área do quadrado = $\pi/4$.

Determine a relação média entre o número de pontos escolhidos e o número de dígitos corretos no valor calculado de π. Estime o número de pontos necessários para determinar corretamente o valor π até a 50^{a} casa decimal.

5.12 Para converter a saída de um neurônio de uma função degrau para uma função sigmóide (ver página 268), pode-se substituir uma afirmação da forma

"Deixe que X seja a soma ponderada das entradas; então a saída é 1 se $X > 0$, senão ela é 0" pela afirmação "Deixe que X seja a soma ponderada das entradas; então a saída é $1/(1\ (4x) + (4x)\ e^{-X})$". (a) Verifique que, quando $X \rightarrow -\infty$, $1/(1\ (4x) + (4x)\ e^{-X}) \rightarrow 0$, quando $X \rightarrow +\infty$, $1/(1\ (4x) + (4x)\ e^{-X}) \rightarrow 1$, e se $X = 0$, $1/(1\ (4x) + (4x)\ e^{-X}) = 0{,}5$. (b) Suponha que a rede para determinar se um ponto se localiza dentro de um triângulo (página 267, embaixo) é alterada desta maneira, de forma que a saída de cada neurônio é descrita pela função sigmóide $1/(1\ (2x) + (2x)\ e^{-X})$, ao invés de pela função degrau, e que um ponto é considerado dentro da área aceitada se a saída da rede for > 0,5. Escreva um programa PERL para determinar que área é definida.

Weblemas

5.1 A bactéria *Pseudomonas fluorescens* e o fungo *Curvularia inaequalis* possuem, cada um, uma cloroperoxidase, uma enzima que catalisa reações de halogenação. Essas enzimas possuem o mesmo padrão de enovelamento?

5.2 Confira a predição dos segmentos de hélice transmembrana na bacteriorodopsina a partir das atribuições de estrutura secundária na estrutura experimental no Protein Data Bank.

5.3 A Gravura VI mostrou a estrutura de um domínio de ligação à tiamina, identificado por M. Gerstein como um dos cinco tipos de padrões de enovelamentos mais comuns que aparecem em Archaea, Bacteria e Eukarya. Utilizando o SCOP, desenhe figuras das outras quatro estruturas.

5.4 Utilizando o resultado do *Weblema* 5.3, ou as figuras disponíveis no livro *Introduction to Protein Architecture: The Structural Biology of Proteins*, represente diagramas simplificados de topologias análogos ao apresentado na Gravura VI para as outras quatro estruturas.

5.5 O gene globina θ_1 humano codifica uma globina ativa? Ou ele é realmente um pseudogene? Envie a seqüência de aminoácidos da globina θ_1 humana para o SWISS-MODEL, incluindo um pedido de relatório WhatCheck no resultado. O que você pode concluir a partir dos resultados sobre a condição da globina θ_1 humana?

5.6 Compare o número de entradas no SCOP, em diferentes categorias, listadas na página 260, com o número de entradas que o SCOP contém atualmente.

5.7 Alinhe as seqüências da quimotripsina γ e da toxina epidermolítica A de *S. aureus* utilizando métodos de alinhamento par a par. Compare o resultado com o alinhamento estrutural mostrado no texto.

5.8 Alinhe a seqüência da elastase de neutrófilo humana e da elastase de *C. elegans*. (a) No alinhamento ótimo, quantos resíduos idênticos existem? (b) Seria plausível construir um modelo para a elastase de *C. elegans* a partir da estrutura da elastase de neutrófilo humana?

5.9 S. Chakravarty e K. K. Kannan resolveram a estrutura da anidrase carbônica com o ligante benzenossulfonamida [Código PDB: 1CZM]. Desenhe figuras do sítio de ligação mostrando a natureza das interações entre a proteína e os ligantes. Descreva a natureza das interações em termos das conclusões obtidas da análise por QSAR.

CAPÍTULO 6

Proteômica e Biologia de Sistemas

Sumário do capítulo

Microarranjos de DNA 313
Análise dos dados de microarranjos 315
Espectrometria de massa 321
Identificação dos componentes de uma mistura complexa 321
Seqüenciamento de proteínas por espectrometria de massa 324
Análise de seqüências de genomas por espectrometria de massa 326
Biologia de sistemas 331
Redes e grafos 333
Estrutura e dinâmica de redes 338
Complexos e agregados de proteínas 340
Propriedades de complexos proteína-proteína 341
Redes de interações protéicas 344
Redes reguladoras 349
Estruturas de redes reguladoras 350
Biologia estrutural de redes reguladoras 356
Leituras recomendadas 359
Exercícios, problemas e weblemas 359

Objetivos do aprendizado

1. Compreender os objetivos da proteômica – a medição de quantidades e de distribuições de proteínas em uma célula ou organismo.
2. Familiarizar-se com os dados derivados de microarranjos e sua aplicação na inferência e interpretação das similaridades e diferenças em padrões de expressão gênica. Compreender as relações entre os dados "brutos" de um microarranjo típico (ver Gravura VII) e a tabela de expressão gênica.
3. Familiarizar-se com os dados derivados de espectroscopia de massa e sua aplicação na análise de misturas de proteínas, no seqüenciamento parcial de proteínas, no seqüenciamento rápido e em grande escala de ácidos nucléicos e na busca por variantes de seqüências genéticas.
4. Apreciar a tendência para uma nova perspectiva: o lema da biologia de sistemas é a integração.
5. Reconhecer a diferença entre redes físicas e redes lógicas.
6. Compreender as características gerais de grafos, incluindo as diferenças entre grafos orientados, não-orientados e rotulados. Compreender a representação de redes por grafos.
7. Apreciar os diferentes tipos possíveis de estados dinâmicos de redes.
8. Compreender as características dos complexos proteína-proteína e proteína-ácido nucléico.
9. Compreender a estrutura de alguns dos pilares das redes reguladoras.

A proteômica é o estudo da distribuição e das interações de proteínas no tempo e no espaço, em uma célula ou um organismo. Métodos experimentais rápidos e em grande escala de análises de dados, incluindo análises de microarranjos e espectroscopia de massa, estão nos fornecendo uma visão em larga escala do sistema de organização de proteínas em organismos vivos e da variação de genes e de proteínas em populações.

O objetivo da biologia de sistemas é combinar proteômica, genômica e outros dados em um quadro integrado da estrutura, dinâmica, logística e, por fim, da lógica funcional dos organismos vivos. Para qualquer proteína de uma célula, a biologia de sistemas irá combinar os estudos acerca da proteína, seu gene, as moléculas que controlam a sua expressão ou sua atividade após ser expressa, e o conjunto de outras proteínas com as quais ela interage. A biologia de sistemas irá organizar, em redes metabólicas, as reações químicas catalisadas pelas enzimas de uma célula e organizar, em redes de controle, os mecanismos que regulam as suas atividades e expressão.

Microarranjos de DNA

Microarranjos de DNA são empregados (1) na análise do conteúdo de mRNA de uma célula para revelar os padrões de expressão de proteínas ou (2) na análise de DNA genômico, para revelar genes ausentes ou mutados.

Ver Quadro Aplicações de microarranjos de DNA, Capítulo 1, página 73

1. **Para uma caracterização integrada de atividades celulares, esperamos determinar quais proteínas estão presentes, onde e em que quantidades.** Para determinar o padrão de expressão dos genes de uma célula, medimos as quantidades relativas de muitos mRNAs diferentes. A hibridização é uma maneira acurada e sensível de detectar a presença de seqüências específicas de ácidos nucléicos. A solução para a análise rápida e em larga escala é realizar diversos experimentos de hibridização em paralelo.
2. **O conhecimento da seqüência do genoma humano pode auxiliar na identificação de genes associados com propensões a doenças.** Algumas doenças (como a fibrose cística) se originam de mutações em um único gene. Nestes casos, isolar uma região pelo método clássico de mapeamento genético pode levar à identificação da lesão. Outras doenças, como a asma, dependem de interações entre muitos genes, com fatores ambientais como complicadores. Para compreender a etiologia de doenças multifatoriais, é necessário determinar e analisar múltiplos padrões de expressão de genes, que podem estar distribuídos em diferentes cromossomos.

Microarranjos de DNA, ou *chips* de DNA, são dispositivos capazes de verificar, em uma amostra, a presença *simultânea* de muitas seqüências.

A idéia básica é: para verificar se um oligonucleotídeo possui uma seqüência particular conhecida, testa-se a sua capacidade de se ligar à seqüência complementar ("um-contra-um"). Para se detectar a presença ou ausência de um oligo de interesse em uma mistura, separa-se a mistura e testa-se cada componente da mistura quanto à afinidade de ligação a um oligo complementar à seqüência de interesse ("muitos-contra-um"). Essas técnicas são denominadas Northern ou Southern *blot*. Para detectar a presença ou ausência de *muitos* oligos em uma mistura, deve-se sintetizar um conjunto de oligos, complementar a cada seqüência da lista de interesse, e testar cada componente da mistura para ligação a cada membro do conjunto de oligos complementares ("muitos-contra-muitos"). Microarranjos constituem um método em larga escala rápido e eficaz para a realização destes testes em paralelo. Eles também permitem a medição dos níveis de expressão de milhares ou até dezenas de milhares de genes em uma única amostra.

Para realizar análises de hibridização em paralelo, uma grande quantidade de oligômeros de DNA é fixada em pontos determinados de um suporte rígido, em um arranjo bidimensional regular. A mistura a ser analisada é preparada com marcadores fluorescentes para permitir a detecção dos híbridos. Após a exposição do arranjo à mistura, cada componente do arranjo, ao qual algum componente da mistura se ligou, possui um marcador. Como conhecemos a seqüência de cada oligômero sonda em cada um dos pontos de ligação do arranjo, a determinação das *posições* das sondas identifica suas seqüências. Esta metodologia analisa os componentes presentes na amostra.

Microarranjos de DNA, como o descrito, são feitos de pequenas lâminas (*chips*) de vidro ou náilon, geralmente de 2 cm^2. Os oligonucleotídeos são fixados ao *chip* em um arranjo quadrado, com densidades que variam entre 10.000 e 250.000 posições por cm^2. O ponto de ligação pode ser tão diminuto quanto ~150 μm de diâmetro. A malha geralmente possui uns poucos centímetros de extensão. Um *chip de le-*

vedura contém mais de 6.000 oligos, correspondentes a todos os genes conhecidos de *Saccharomyces cerevisiae*. Um arranjo de DNA, ou um *chip* de DNA, pode conter 400.000 oligômeros sonda. Note que este número é maior do que o número total de genes em organismos superiores (excluindo-se os genes de imunoglobulinas).

Para analisar a mistura, é necessário expô-la ao microarranjo sob condições que favoreçam a hibridização e depois lavar a fração da amostra que não se ligou às sondas. Para comparar dois conjuntos de oligos, marcam-se as amostras com fluoróforos de cores diferentes (Gravura VII). Com o auxílio de um escâner, os dados do arranjo são coletados em um formato lido por computadores.

Existem diferentes tipos de *chips*, construídos de acordo com a investigação que se deseja realizar e com diferentes tipos de DNA imobilizados.

1. Em um ***chip* de expressão**, os oligos imobilizados são amostras de cDNA, tipicamente entre 20 e 80 pb, derivados de mRNAs de genes conhecidos. O objetivo deste tipo de experimento é determinar os padrões de expressão de genes que correspondam às amostras de cDNA. Esta é, sem dúvida, a aplicação mais comum dos microarranjos. A amostra de interesse pode ser uma mistura de mRNAs de tecido normal ou afetado por alguma doença.
2. Em um **microarranjo de análise de mutações**, procura-se por padrões de polimorfismos de um único nucleotídeo (SNPs).
3. Na **hibridização genômica**, procura-se pela perda ou pelo ganho de genes, ou por alterações no número de cópias. As seqüências sonda, fixadas no *chip*, são grandes segmentos de DNA genômico obtidos de porções conhecidas dos cromossomos, com comprimentos entre 500 e 5.000 pb. As misturas sonda contêm DNA de um estado normal ou afetado por alguma doença. Por exemplo, alguns tipos de câncer se originam da deleção de cromossomos, o que pode ser identificado com os microarranjos.

Os microarranjos são capazes de comparar concentrações de componentes da amostra. Isso permite uma investigação quantitativa de respostas a alterações de condições. No entanto, a precisão é baixa. Além disso, os níveis de mRNA detectados pelo arranjo nem sempre refletem, corretamente, os níveis de proteína. De fato, em geral é feita uma transcrição reversa do mRNA em cDNAs mais estáveis para a análise por microarranjos, e o rendimento desta etapa nem sempre é uniforme. Assim, os dados dos microarranjos são semiquantitativos, no sentido de que uma distinção entre presença e ausência é possível, mas a determinação de níveis relativos de expressão em um experimento controlado é mais difícil e a medição de níveis de expressão absolutos está além da capacidade atual das técnicas de microarranjo.

Recursos na *web*: Bancos de dados de microarranjos

Os microarranjos fornecem mais uma maneira de produção de dados rápida e em grande escala na bioinformática. Um padrão chamado MIAME (Minimum Information About a Microarray Experiment) descreve os conteúdos e o formato da informação a ser coletada em um experimento e depositada em bancos de dados.

Os principais bancos de dados públicos e disponíveis de microarranjos incluem:

O EBI mantém um banco de dados, ArrayExpress:

`http://www.ebi.ac.uk/arrayexpress/`

O NCBI hospeda o banco de dados Gene Expression Omnibus:

`http://www.ncbi.nlm.nih.gov/geo/`

O Stanford Microarray Database (SMD):
`http://genome-www5.stanford.edu/MicroArray/SMD/`

Uma lista de bancos de dados de microarranjos para plantas se encontra em:
`http://www.univ-mont2.fr/~plant_arrays/databases.html`

Para listas adicionais, consulte as referências seguintes:

Butte, A. (2002), The use and analysis of microarray data, *Nat. Revs. Drug Discov.*, **1**, 951-960.

Penkett, C. J. & Bähler, J. (2004), Getting the most from public microarray data, *European Pharmaceutical Review*, **1**, 8-17.

Análise dos dados de microarranjos

Os dados brutos de um experimento de microarranjo são uma imagem na qual a intensidade e a cor da fluorescência refletem o grau de hibridização com sondas alternativas. Os dois conjuntos de sondas são marcados com fluoróforos vermelhos e verdes. Se apenas uma das sondas hibridizar, aparecerão pontos vermelhos na placa; se apenas a outra sonda hibridizar, aparecerão pontos verdes. Caso as duas sondas hibridizem, os pontos resultantes serão amarelos (vermelho + verde, ver Gravura VII).

O objetivo inicial do processamento dos dados é obter uma **tabela de expressão gênica**. Essa tabela é uma matriz na qual as linhas correspondem a genes diferentes e as colunas correspondem a diferentes amostras. Pontos distintos em um padrão de microarranjo, como o mostrado na Gravura VII, correspondem a genes diferentes. Para cada gene, resultados de diferentes conjuntos de amostras aparecem em vermelho ou verde (ou ambos ou nenhum). Existe uma grande redundância dos oligos em um microarranjo – cada gene pode ser representado por vários pontos, correspondendo a regiões diferentes da seqüência do gene. A inclusão de controles com malpareamentos deliberados permite a verificação dos dados. Tipicamente, um gene pode corresponder a ~30-40 pontos.

As amostras podem variar de acordo com as condições experimentais e/ou estados fisiológicos, ou podem ser extraídas de indivíduos diferentes, tecidos diferentes, ou estágios de desenvolvimento diferentes.

O processo de redução de dados para produzir a matriz de expressão gênica envolve muitos detalhes técnicos de processamento de imagens, checagem de controles internos, considerações de dados ausentes, seleção das medidas confiáveis e padronização dos resultados de diferentes arranjos em escalas consistentes. A tabela de expressão gênica derivada indica valores *relativos* de níveis de expressão. Uma mudança nos níveis de expressão de um gene entre duas amostras, por um fator igual ou maior do que 1,5-2,0, normalmente é considerada significativa.

A extração de informações biológicas confiáveis da tabela de expressão gênica não é trivial. Apesar do uso extensivo de controles internos, existe um nível considerável de ruído na técnica experimental. Em muitos casos, a variabilidade é inerente às próprias amostras. Microrganismos podem ser clonados e animais podem sofrer endocruzamento em um grau considerável de homogeneidade. Porém, os experimentos realizados com RNA obtido de amostras humanas – por exemplo, um conjunto de pacientes que sofre de uma doença e um conjunto correspondente de controles saudáveis – estão à mercê da ampla variação individual que os seres humanos apresentam. De fato, animais obtidos

por endocruzamento, e mesmo amostras de culturas de tecidos eucarióticos aparentemente idênticos, apresentam grande variabilidade.

Outra desvantagem intrínseca – e bastante séria – na interpretação dos dados de expressão gênica é o fato de o número de genes ser muito maior do que o número de amostras. Computacionalmente, estamos tentando entender a relação entre um espaço de muitas variáveis (os genes) com um espaço de observações (o fenótipo), a partir da medição de apenas alguns pontos (as amostras). A dispersão das observações não confere uma cobertura adequada dos dados. Métodos estatísticos têm uma grande responsabilidade na análise para nos dar confiança na significância de nossas conclusões.

Duas abordagens gerais para a análise de uma matriz de expressão gênica envolvem (a) *comparações focalizadas nos genes*, ou seja, comparações de distribuições dos padrões de expressão de diferentes genes pela comparação das linhas em uma matriz de expressão ou (b) *comparações focalizadas nas amostras*, ou seja, comparações de perfis de expressão de diferentes amostras pela comparação das colunas da matriz de expressão:

(a) **Comparações focalizadas nos genes: Como os padrões de expressão gênica variam entre diferentes amostras?** Suponha que se saiba que um gene está relacionado com uma doença, ou com alguma alteração das condições fisiológicas em resposta a alterações do ambiente. Outros genes que são co-expressos com o gene conhecido podem participar de processos que contribuam para a doença ou para a alteração das condições fisiológicas. De forma geral, se duas linhas (dois genes) na matriz de expressão gênica apresentam padrões similares de expressão nas amostras, isso sugere um padrão de regulação comum, e possivelmente alguma relação entre as suas funções, incluindo, mas não se limitando a, uma possível interação física.

(b) **Comparações focalizadas nas amostras: Como as amostras diferem nos seus padrões de expressão gênica?** Um conjunto de diferenças consistente entre as amostras pode caracterizar as classes de origem das amostras. Se as amostras são oriundas de diferentes grupos-controle (por exemplo, animais doentes e animais saudáveis), as amostras de grupos diferentes apresentam diferenças consistentes nos seus padrões de expressão gênica? E se apresentarem essas diferenças, dada uma nova amostra, é possível classificá-la adequadamente de acordo com o seu padrão de expressão gênica observado?

Como, então, podemos medir a similaridade entre diferentes linhas ou colunas? Cada linha ou coluna da matriz de expressão pode ser considerada um vetor em um espaço de muitas dimensões. Os vetores linha (uma linha corresponde a um gene), onde cada entrada se refere ao mesmo gene em diferentes amostras, possuem tantos elementos quantos forem as amostras. Os vetores coluna (uma coluna corresponde a uma amostra), onde cada entrada se refere a um gene diferente em uma amostra individual, possuem tantos elementos quantos forem os genes presentes na amostra. É possível calcular o "ângulo" entre diferentes vetores linha ou entre diferentes vetores coluna para fornecer a medida de suas similaridades. Portanto, é natural indagar se subconjuntos de pontos formam grupos naturais – pontos com alta similaridade recíproca – que caracterizam ou conjuntos de genes ou conjuntos de amostras.

Dependendo da origem das amostras, do que já se sabe sobre elas e do que se deseja descobrir sobre elas, a análise dos dados pode prosseguir de diferentes maneiras.

(1) O caso mais simples é um estudo cuidadosamente controlado, utilizando dois conjuntos de amostras de *características conhecidas*. Por exemplo, as

amostras podem ser obtidas a partir de bactérias crescidas na presença ou ausência de um fármaco, a partir de moscas jovens e moscas adultas, ou de pacientes humanos saudáveis e portadores de alguma doença. Podemos focalizar a investigação na seguinte questão: que diferenças no padrão de expressão gênica caracterizam estes dois estados? Podemos desenvolver uma regra de classificação tal que, dada uma nova amostra, podemos atribuí-la corretamente à sua própria classe? Esta metodologia poderia ser aplicada no diagnóstico de doenças. Por exemplo, a determinação do subtipo de leucemia permite um tratamento mais acurado e melhor prognóstico. Havendo a disponibilidade de dados adequados, essa metodologia pode ser estendida para sistemas de mais de duas classes.

Computacionalmente, o treinamento de um algoritmo de classificação, como sugerido acima, é denominado "aprendizado supervisionado". O padrão de expressão de cada amostra é dado por um vetor correspondente a uma única coluna da matriz. Esse vetor corresponde a um ponto em um espaço multidimensional – com tantas dimensões quantos forem os genes. Nos casos favoráveis, os pontos estarão distribuídos em regiões diferentes do espaço. Assim, um cientista, ou um programa de computador, será capaz de demarcar o limite entre as amostras. Em outros casos, a separação das classes será mais difícil. Considere a distribuição de jogadores de futebol durante uma partida. No início do jogo, a linha de meio de campo separa os dois times, ou seja, a linha de meio de campo o divide em duas regiões, e cada região contém exclusivamente jogadores de um dos times. Durante a partida, os times se misturam e é muito difícil determinar os limites entre cada um deles.

(2) Em uma situação experimental diferente, podemos não ser capazes de *predeterminar* as diferenças entre as categorias de amostras. Ao contrário, esperamos extrair a classificação das amostras da análise realizada. O objetivo é agrupar os dados para *identificar* as classes de amostras e as diferenças entre os genes que as caracterizam.

Muitos algoritmos de agrupamento têm sido aplicados aos dados derivados de microarranjos, incluindo aqueles que tentam lidar simultaneamente com o *número* de grupos e as *fronteiras* entre eles. Todos os algoritmos precisam lidar com a dificuldade que surge da escassez de amostragem, devido ao caráter multidimensional do espaço de medidas. Algumas vezes é possível simplificar o problema identificando-se um número pequeno de combinações de genes que representem uma grande porção da variabilidade. Esta abordagem é chamada de **redução de dimensionalidade** (ver Quadro).

Redução de dimensionalidade

A distribuição dos dados de expressão gênica em um espaço com elevado número de dimensões significa que (1) a cobertura do espaço com um número limitado de amostras é esparsa e (2) é difícil visualizar a distribuição dos pontos da amostra. Em alguns casos, a distribuição pode depender principalmente das poucas variáveis equivalentes, sendo bastante vantajoso encontrá-las e transformar os dados de acordo.

Um exemplo simples ilustra a idéia básica. Considere uma distribuição de grupos de pessoas fazendo piqueniques em uma praia. Represente a

→

Redução de dimensionalidade (*continuação*)

posição de cada pessoa pelas coordenadas x, y e z da ponta de seus narizes. Considere o eixo x paralelo à linha da praia, o eixo y perpendicular à linha da praia e o eixo z vertical. Obviamente a altura é irrelevante: esta é, sem dúvida, uma distribuição bidimensional e não tridimensional. Para classificar as pessoas nos grupos (talvez famílias ou grupos de surfistas), os eixos x e y comportam todos os dados significativos e a coordenada z contém apenas dados irrelevantes, como a altura das pessoas e se elas estão em pé ou sentadas na areia. Neste caso, para reduzir a dimensionalidade de três para duas, precisamos apenas ignorar a coordenada z. (Naturalmente, se a maré subir e se a área da praia se tornar menor, a dimensão ao longo da linha da praia fornecerá toda a informação necessária e a dimensionalidade poderia ser reduzida de duas para uma.)

Agora suponha que as pessoas não estejam distribuídas pela praia, mas escalando uma rocha vertical paralela à linha da praia. Esta também é uma distribuição bidimensional e não tridimensional, mas neste caso as coordenadas x e z é que possuem as informações relevantes.

Em casos mais complexos, a redução de dimensionalidade exige mais do que a simples escolha de qual coordenada ignorar. Suponha que as pessoas estejam distribuídas em uma rampa de esqui. Para reduzir a distribuição de três para duas dimensões, não basta simplesmente ignorar uma coordenada, mas, sim, *projetar* os dados no plano oblíquo paralelo à rampa. (O plano paralelo ao piso da rampa é oblíquo a um sistema de coordenadas orientado ao longo das direções vertical e horizontal.) A idéia de uma *projeção* dos dados em um espaço de dimensões menores, o qual contém os componentes importantes da variação, é o ponto principal do método.

Os problemas práticos da análise de dados são mais difíceis do que estes exemplos simples. Primeiro, as dimensões iniciais são muito mais do que três e a redução de dimensionalidade é potencialmente muito maior. Segundo, o modo como a redução é realizada não é trivial, pois não temos uma visualização fácil do espaço físico e da distribuição das pessoas na praia, ou na superfície da rocha, ou na rampa de esqui.

No entanto, as questões a serem respondidas permanecem: ao longo de quais direções deveríamos projetar os dados de forma que a maior discriminação seja mantida utilizando o menor número de dimensões? Métodos matemáticos conhecidos como **Análise da Componente Principal** (PCA) usando a **Decomposição de Valores Únicos** (SVD) podem resolver este problema. Esses métodos selecionam automaticamente um novo sistema de coordenadas que melhor represente a variabilidade dos dados ao longo do menor número de eixos e, para cada novo eixo de coordenadas, o cálculo fornece uma medida da contribuição desta coordenada que represente a variabilidade dos dados.

Apesar de duas dimensões poderem não conter todos os componentes importantes de uma variação, sempre podemos escolher as duas melhores projeções bidimensionais para a representação gráfica do resultado, o que possui a imensa vantagem de permitir a visualização e a avaliação dos dados pelos cientistas. (Distribuições tridimensionais também podem ser representadas visualmente, através de gráficos, mas com um grau maior de dificuldade.)

Estudo de Caso 6.1: Interpretação de dados de um microarranjo: Regulação de genes por BRCA1 e implicações para o papel da disfunção ou silencialidade de BRCA1 na carcinogênese

O gene *BRCA1* codifica um supressor de tumor. Ele se encontra mutado em aproximadamente 50% dos pacientes com predisposição familiar ao câncer de mama e ovário. Uma única cópia do alelo defeituoso do *BRCA1* é suficiente para aumentar o risco, pois, em qualquer célula a cópia normal do gene pode ser perdida ou, em uma pequena porcentagem dos casos, ser inativada por metilação do promotor.

A proteína BRCA1 é composta por 1.863 resíduos de aminoácidos. Ela possui um domínio *ring finger* na porção N-terminal, seguido de uma região helicoidal *coiled-coil* e por dois domínios BRCT em *tandem*, os quais são sítios de ligação para outras proteínas, além de atuarem na regulação da transcrição (BRCT é a abreviação para domínio BRCA C-terminal).

A proteína BRCA1 interage com muitas outras proteínas, formando complexos funcionais envolvidos em várias atividades diferentes, incluindo:

- **Percepção e sinalização de lesões no DNA.** O gene *BRCA1* responde a diversos tipos de dano ao DNA – por exemplo, quebras na dupla hélice – e ativa os mecanismos de reparo apropriados para cada um.
- **Preservação da estrutura do cromossomo.** Como a integridade dos cromossomos pode ser afetada como uma *conseqüência* do reparo impreciso dos danos na cadeia de DNA, estas funções estão relacionadas.
- **Mediação dos testes de verificação em pontos específicos do ciclo celular**, em parte pela regulação da transcrição dos genes que codificam proteínas envolvidas nas atividades de verificação.

Um consenso sobre o *BRCA1* é que a proteína codificada medeia respostas ao dano no DNA pela ativação de mecanismos de reparo e, no caso de este reparo não ser bem-sucedido, de mecanismos de verificação e controle que impedem que as células com danos não-reparados se propaguem. A perda das funções da proteína BRCA1 leva ao acúmulo de danos no DNA celular, aumentando as chances de transição para um estado cancerígeno.

A variedade e a complexidade dos processos envolvendo BRCA1 tornam difícil a compreensão do mecanismo detalhado da sua relação com o câncer:

1. A formação de tumores é uma conseqüência direta da perda de uma ou mais funções da BRCA1 e das proteínas com que interage? Se for, quais são essas funções ou interações?
2. Qual é a importância da regulação transcricional do gene *BRCA1* pelo produto de outros genes ou de outros genes, pela BRCA1, ou ambos? Até que ponto alterações no padrão de expressão envolvendo *BRCA1* levam *indiretamente* à tumorigênese? Perceberemos que a distinção entre os efeitos diretos e indiretos não é muito bem definida: BRCA1 se liga diretamente a algumas proteínas cuja expressão ela regula.
3. Mecanismos de reparo do DNA são comuns em muitos tipos celulares. Por que a disfunção ou silencialidade do gene *BRCA1* leva ao maior risco de câncer de mama e ovário (e de outros tecidos epiteliais, como pâncreas e próstata)?

→

Estudo de Caso 6.1 (*continuação*)

Uma função da BRCA1 é o controle da transcrição. Para estudar o contexto regulador da relação entre *BRCA1* e o risco de câncer, Welcsh e colaboradores utilizaram análises de microarranjos para comparar o padrão de expressão de genes em células que expressavam níveis elevados e reduzidos de BRCA1, empregando uma linhagem celular em que a expressão de BRCA1 foi induzida seletivamente (ver Gravura VIII). O *chip* utilizado para a detecção da resposta continha oligonucleotídeos representando ~6.800 genes humanos. (Note que esta é uma fração relativamente pequena do proteoma humano total.)

Os resultados apontaram 373 genes com expressão diferenciada, e em quantidades significativas e reproduzíveis, em resposta a níveis mais elevados de expressão de BRCA1. Dentre esses genes, 57 apresentaram aumento da expressão e 15 apresentaram uma diminuição da expressão, para os quais os níveis de expressão foram alterados por fatores ≥ 2. Estes candidatos a envolvimento com as funções de BRCA1, relevantes para a tumorigênese, foram avaliados quanto à expressão diferencial em tecidos de pacientes com câncer e controles saudáveis.

O agrupamento da matriz de expressão gênica mostra a distinção clara entre os genes com aumento e com diminuição da expressão e dá uma idéia da variabilidade entre as réplicas (Gravura VIII). Muitas das proteínas codificadas pelos genes que apresentam aumento de expressão são receptores hormonais e proteínas estruturais. Muitas das proteínas codificadas pelos genes com diminuição da expressão estão envolvidas na replicação e tradução do DNA.

Dentre os genes identificados no estudo, destacam-se os seguintes:

Consistente com o surgimento tecido-específico de tumores, como resultado da disfunção do gene BRCA1, alguns dos genes com padrões de expressão alterados estão envolvidos com vias de controle mediadas por estrógeno, sugerindo uma possível resposta para o enigma da especificidade. O conjunto de proteínas implicado inclui a ciclina D1 e myc, que têm aumento da expressão em concentrações mais baixas de *BRCA1*. O aumento da expressão da ciclina D1 e de myc é observado em 20% dos casos de câncer de mama, o que está de acordo com a repressão destas proteínas pela BRCA1 funcional. (Para comparação com os dados clínicos, níveis baixos de expressão de BRCA1 estão relacionados com pacientes que apresentam as funções de BRCA1 reduzidas ou ausentes, ou seja, grupo de alto risco; níveis elevados de expressão de BRCA1 são análogos aos controles normais. Contudo, os experimentos de Welcsh e colaboradores não tentaram reproduzir os níveis de expressão endógena reais de BRCA1 observados nos pacientes e nos seus controles normais.)

Por outro lado, as proteínas JAK e STAT têm uma diminuição de expressão em baixas concentrações de *BRCA1*. Essas proteínas estão envolvidas como inibidores de crescimento nas vias de controle que regulam a proliferação, diferenciação, apoptose e transformação. A perda da atividade BRCA1 reduziria os níveis de JAK1 e STAT1, promovendo a proliferação celular e a redução da apoptose. Estes dados estão de acordo com a observação de que camundongos que não expressam *Stat1* desenvolvem tumores mais rapidamente do que camundongos normais.

As relações detectadas por Welcsh e colaboradores são parte da rede de controle celular. No entanto, sabe-se que alguns dos produtos dos genes

regulados por BRCA1 também estão envolvidos na formação de complexos funcionais com BRCA1. Por exemplo, o produto do gene *myc* – um oncogene potente – liga-se à BRCA1, sugerindo uma inibição direta da função de myc por BRCA1. Dessa forma, a redução dos níveis de BRCA1 terá dois efeitos: a redução da inibição de *myc* via ligação direta e o aumento da sua expressão pela perda da repressão transcricional.

Portanto, *myc* está relacionado com BRCA1 tanto por interações físicas quanto reguladoras. Veremos nas seções posteriores deste capítulo que a idéia de duas redes paralelas de interações nas células – interações físicas e reguladoras – são distinções fascinantes. Porém, é uma distinção difícil de manter em um sistema, tal como a função de BRCA1, no qual as duas interações estão intimamente conectadas.

Espectrometria de massa

A espectrometria de massa é uma técnica física que caracteriza moléculas pelas medições das massas de seus íons. Estudos em larga escala de padrões de expressão de proteínas exigem métodos que forneçam grandes quantidades de dados, de forma rápida, e que tenham alta precisão. A espectrometria de massa preenche essas condições, o que estimulou o seu desenvolvimento em uma tecnologia madura e amplamente utilizada. Suas aplicações na biologia molecular incluem:

- Identificação rápida dos componentes de uma mistura complexa de proteínas.
- Seqüenciamento de proteínas e ácidos nucléicos, incluindo o seqüenciamento genômico rápido e em grande escala e estudos populacionais sobre variabilidade genética.
- Análises de modificações pós-traducionais ou substituições em relação a uma seqüência esperada.

Identificação dos componentes de uma mistura complexa

Inicialmente, os componentes são separados por eletroforese, as proteínas isoladas são digeridas por tripsina para originar fragmentos de peptídeos com r.m.m (*relative molecular mass* – massa molecular relativa) entre 800 e 4.000 (ver Figura 6.1). A tripsina cliva proteínas após resíduos de Lis e Arg. Em uma composição típica de aminoácidos, uma proteína com 500 resíduos de aminoácidos irá gerar cerca de 50 fragmentos pela digestão com tripsina. O espectrômetro mede as massas dos fragmentos com alto grau de precisão. A lista das massas dos fragmentos, chamada de **impressão digital das massas de peptídeos**, caracteriza a proteína (ver Figura 6.2). A consulta em bancos de dados de massas de fragmentos identifica as amostras desconhecidas.

A construção de bancos de dados de massas de fragmentos é um cálculo simples a partir da seqüência de aminoácidos de proteínas conhecidas ou da tradução de ORFs de genomas, ou ainda (em uma emergência) de segmentos de bibliotecas de ESTs. Os fragmentos correspondem a segmentos clivados pela tripsina em resíduos de lisina e arginina, e as massas dos aminoácidos são conhecidas. (Note que a tripsina não é capaz de clivar a ligação peptídica entre Lis-Pro e também pode falhar na clivagem da ligação entre Arg-Pro.)

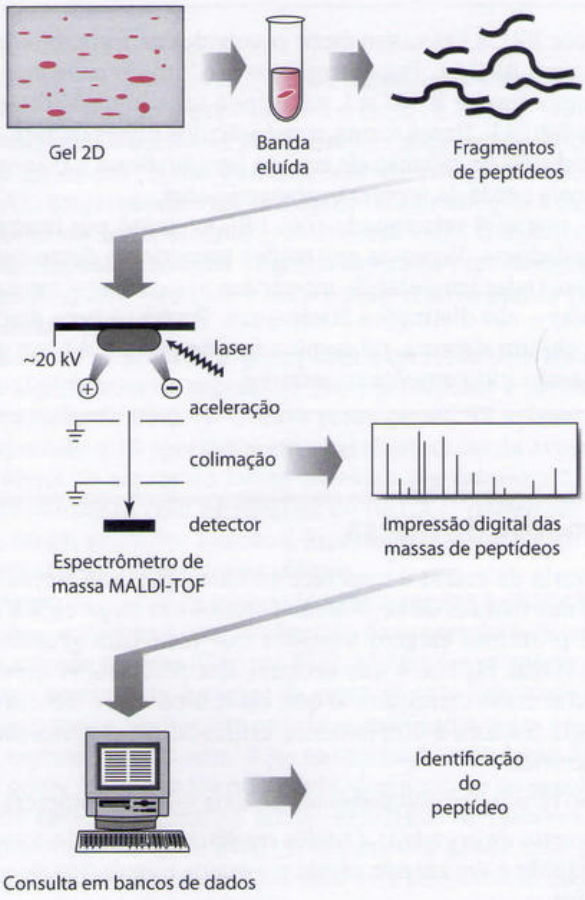


Figura 6.1 Identificação dos componentes de uma mistura de proteínas pela eluição de bandas individuais, digestão e determinação da impressão digital dos fragmentos de peptídeos por espectrometria de massa MALDI-TOF (Matrix-Assisted Laser Desorption Ionization – Time of Flight), seguida pela busca do conjunto das massas dos fragmentos em um banco de dados.

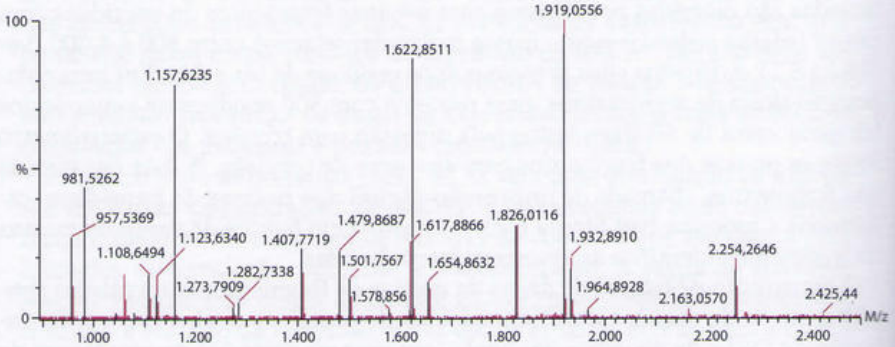


Figura 6.2 Espectro de massa de uma digestão por tripsina. Dos 21 picos mais altos (em preto), 15 apresentam a massa esperada para os peptídeos, digeridos com tripsina, da subunidade de 39 kDa do complexo I mitocondrial de vaca. Isso é o suficiente para uma identificação positiva. (Esta figura é uma cortesia do Dr. I. Fearnley.)

Recursos na *web*: Identificação de proteínas a partir de impressões digitais das massas de peptídeos

Dois bancos de dados utilizados comumente que compilam impressões digitais das massas de peptídeos de proteínas conhecidas:

`www.matrixscience.com`
`prospector.ucsf.edu/prospector/4.0.8/html/msfit.htm`

A espectrometria de massa (EM) é sensível e rápida. As impressões digitais das massas de peptídeos podem identificar proteínas em quantidades inferiores a picomoles. Medidas de massas de fragmentos com uma resolução superior a 0,1 unidades de massa são suficientemente boas para analisar misturas isotópicas. A EM é um método de análise rápida e em grande escala, capaz de processar 100 bandas por dia (embora a preparação das amostras exija um tempo maior). Porém, existem limitações. Apenas proteínas de seqüência conhecida podem ser identificadas pelas impressões digitais de massas de peptídeos, pois somente as massas preditas dos seus fragmentos são incluídas nos bancos de dados. (Como em outros métodos de identificação de impressões digitais, é possível mostrar que duas proteínas de amostras diferentes são provavelmente a mesma, ainda que as suas identificações não sejam possíveis.) As modificações pós-traducionais também interferem na metodologia pela alteração das massas dos fragmentos.

Os resultados apresentados na Figura 6.2 são de um experimento no qual as massas moleculares dos íons foram determinadas a partir do seu tempo de vôo (*time-of-flight*, TOF) sobre uma distância conhecida, como é ilustrado na Figura 6.1. A operação do espectrômetro envolve as seguintes etapas:

1. Produção da amostra na forma ionizada, na fase de vapor.
2. Aceleração dos íons em um campo elétrico. Cada íon emerge com uma velocidade proporcional à razão entre sua carga e massa.
3. Passagem dos íons por uma região livre do campo elétrico onde eles são colimados.
4. Detecção do tempo de chegada dos íons. O tempo de vôo (TOF) indica a razão massa/carga dos íons.
5. O resultado das medidas é um gráfico mostrando o fluxo como uma função da razão massa/carga dos íons detectados.

Como as proteínas são bastante delicadas, o grande desafio é vaporizá-las e ionizá-las sem causar danos. Dois métodos de "ionização suave" que contornam esse problema são:

1. **Ionização/Dessorção de Matriz Assistida por Laser (MALDI – Matrix-Assisted Laser Desorption Ionization)**. A amostra é introduzida seca no espectrômetro, misturada a um substrato ou matriz, que modera a transferência de energia. Um pulso de laser, absorvido inicialmente pela matriz, vaporiza e ioniza a proteína. A combinação MALDI-TOF, que produziu os resultados mostrados na Figura 6.2, é uma configuração experimental bastante comum.
2. **Ionização por Eletropulverização (ESI – Electrospray Ionization)**. Este método inicia com a amostra na forma líquida. A sua evaporação através de um pequeno capilar com um campo elétrico na extremidade cria um aerossol

com gotículas altamente carregadas. Com a evaporação do solvente, as gotículas se contraem, aproximando as cargas e aumentando a força de repulsão entre elas. As gotículas acabam explodindo em gotículas ainda menores, cada qual com carga total reduzida. Esse processo é repetido até a obtenção final de íons, que podem ter múltiplas cargas, livres do solvente. Esses íons são transferidos para uma câmara de alto vácuo do espectrômetro de massa.

Como a amostra está inicialmente na forma líquida, o método ESI permite a sua automação, na qual uma mistura dos peptídeos digeridos com tripsina passa por uma cromatografia líquida de alto desempenho (HPLC – Hight-Performance Liquid Chromatography) diretamente para o espectrômetro de massa.

Seqüenciamento de proteínas por espectrometria de massa

A fragmentação de um peptídeo produz uma mistura de íons. As condições em que a clivagem ocorre, principalmente nas ligações peptídicas, resultam em uma série de íons cujas diferenças de massas equivalem à massa de um único resíduo de aminoácido (Figura 6.3). A seqüência de aminoácidos do peptídeo pode, portanto, ser deduzida da análise do espectro de massas (Figura 6.4), que pode apresentar ambigüidades: por exemplo, Leu e Ile possuem a mesma massa e não podem ser diferenciadas em fragmentos clivados somente nas ligações peptídicas. Discrepâncias em relação às massas dos aminoácidos-padrão indicam modificações pós-traducionais. Na prática, a seqüência de cerca de 5 a 10 aminoácidos pode ser determinada a partir de um peptídeo com tamanho inferior a 20-30 resíduos.

Atualmente, os fragmentos são produzidos *in situ*: primeiro o peptídeo é vaporizado e depois fragmentado por uma dissociação induzida por colisão (CID – Collision-Induced Dissociation) com o gás argônio. Esta metodologia requer dois analisadores de massa operando em seqüência no mesmo instrumento (chamado MS/MS). A amostra vaporizada passa inicialmente por um analisa-

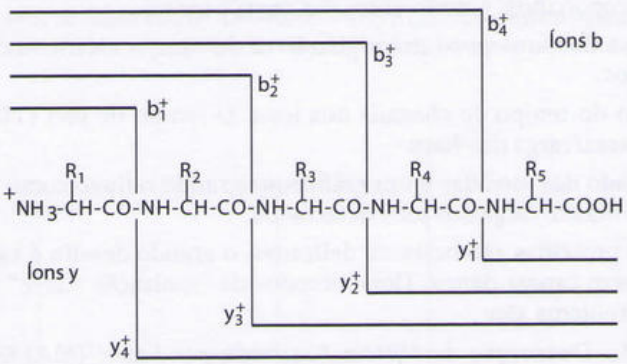


Figura 6.3 Fragmentos produzidos pela clivagem das ligações peptídicas de um peptídeo pequeno. Os íons b contêm a porção N-terminal, e os íons y, a porção C-terminal. A diferença das massas entre íons b sucessivos ou entre íons y sucessivos corresponde à massa de um único resíduo de aminoácido, a partir do qual a seqüência peptídica pode ser determinada. Restam duas ambigüidades: Leu e Ile possuem a mesma massa e não podem normalmente ser diferenciadas; Lis e Gln possuem praticamente a mesma massa e em geral também não podem ser distinguidas. Na dissociação induzida por colisão, a quebra das ligações pode ser limitada em grande parte às ligações peptídicas, mantendo-se as energias de impacto baixas. Colisões de energia mais elevadas podem fragmentar cadeias laterais, úteis para distinguir entre Leu/Ile e Lis/Gln.

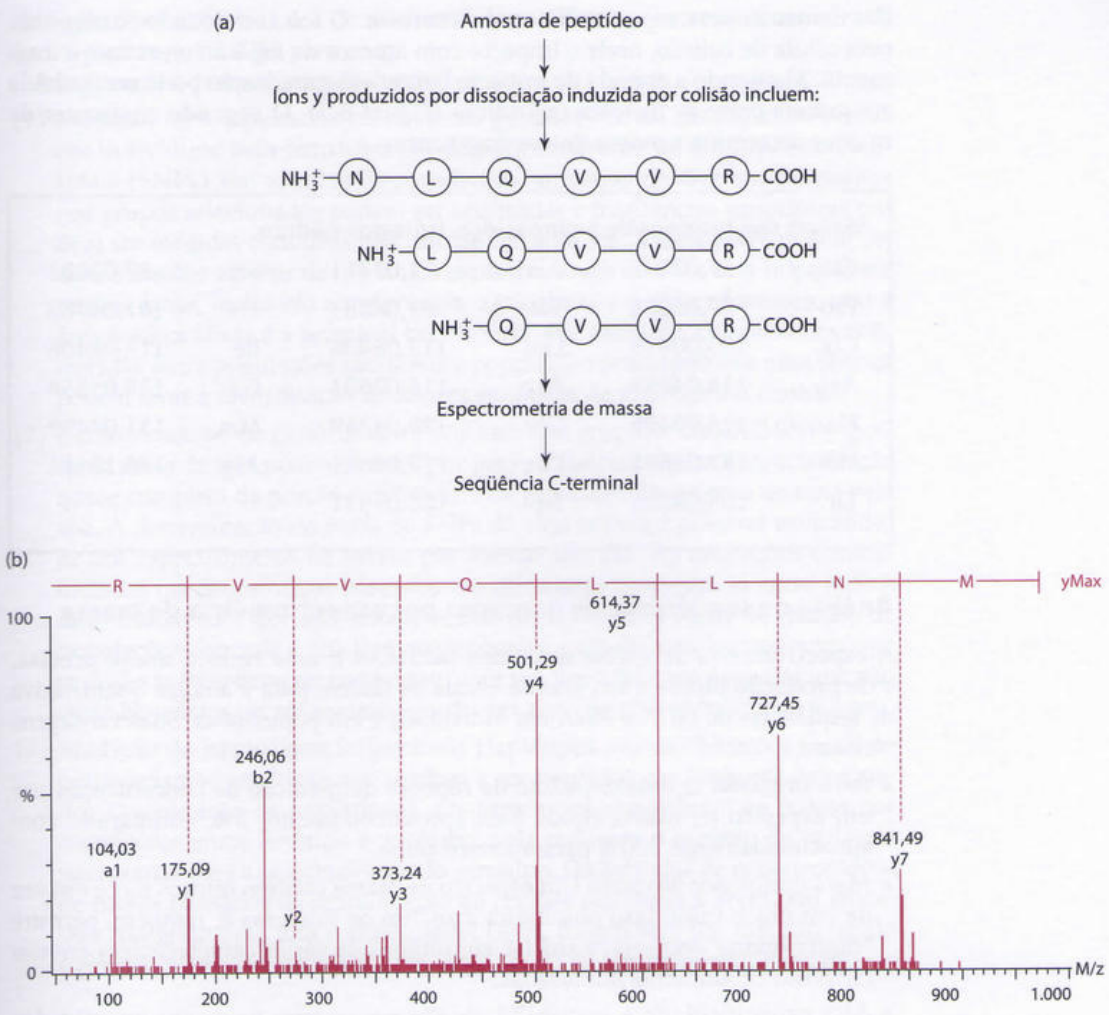


Figura 6.4 Seqüenciamento de peptídeos por espectrometria de massa. A dissociação induzida por colisão (CID – Collision-Induced Dissociation) produz uma mistura de íons. (a) A mistura contém uma série de íons que diferem entre si de acordo com as massas dos aminoácidos sucessivos da seqüência. Na metodologia CID, os íons não são produzidos em seqüência, como é sugerido pela lista. A medida do espectro de massas os arranja em ordem da razão massa/carga automaticamente. (b) Espectro de massa dos fragmentos disponíveis para determinação da seqüência C-terminal. A maior estabilidade dos íons y em relação aos íons b nos fragmentos oriundos da digestão por tripsina simplifica a interpretação do espectro. As diferenças de massa entre picos sucessivos de íons y são iguais às massas dos resíduos individuais de aminoácidos sucessivos na seqüência. Como os íons y contêm a porção C-terminal, o pico de íon y de menor massa equivale ao resíduo C-terminal, etc., e dessa forma a seqüência é obtida "no reverso". Neste experimento, os dois resíduos de leucina desta seqüência não puderam ser distinguidos dos resíduos de isoleucina. Fonte: Carroll, J., Fearnley, I. M., Shannon, R. J., Hirst, J. & Walker, J. E. (2003), Analysis of the subunit composition of complex I from bovine heart mitochondria. *Mol. Cell. Proteomics*, **2**,117-126 (Figura suplementar S138).

dor de massa para separar um íon de interesse. O íon selecionado passa então pela célula de colisão, onde o impacto com átomos de argônio o excita e o fragmenta. Mantendo a energia de impacto baixa, a fragmentação pode ser limitada em grande parte às ligações peptídicas (Figura 6.3). O segundo analisador de massas determina a massa desses fragmentos.

Massas dos resíduos de aminoácidos, isótopos-padrão

Gli	57,02146	Ala	71,03711	Ser	87,03203
Pro	97,05276	Val	99,06881	Tre	101,04768
Cis	103,00919	Leu	113,08406	Ile	113,08406
Asn	114,04293	Asp	115,02694	Gln	128,05858
Lis	128,09496	Glu	129,04259	Met	131,04409
His	137,05891	Phe	147,06841	Arg	156,10111
Tir	163,06333	Trp	186,07931		

Análise de seqüências de genomas por espectrometria de massa

A espectrometria de massa de ácidos nucléicos é uma técnica muito precisa, e de produção rápida e em grande escala de dados, para a análise quantitativa de seqüências de DNA e RNA em indivíduos e em populações. Suas vantagens incluem:

- **Alta precisão** O desvio-padrão de réplicas de medição de concentração por um espectro de massa típico é de aproximadamente 3%, comparado com aproximadamente 200% para microarranjos.
- **Mais dados por amostra** Um espectro de massa contém muitos picos em vez de um único valor. Isso possibilita a análise de misturas e, também, permite "multiplexing", ou seja, a análise simultânea de várias características em um conjunto de amostras misturadas.
- **Alta especificidade e sensibilidade** São necessárias pequenas quantidades de amostras. A amplificação por PCR pode aumentar o volume da amostra, além de haver pouco risco de confundir um contaminante com a amostra real. De fato, é possível determinar seqüências de células individuais ou mesmo de uma única fita de DNA.

Na preparação para a quantificação, as amostras são amplificadas por meio de PCR gene-específico com iniciadores de extensão alelo-específicos, para a produção de oligonucleotídeos de fita simples. Os produtos de PCR são purificados e fixados em uma matriz adequada para a vaporização por MALDI e análise das massas. Nenhuma etapa de hibridização é necessária para a detecção. O agrupamento de várias amostras em um arranjo permite a automação da coleta de dados. (As taxas de produção de dados podem chegar a 10^5 espectros por instrumento por dia.)

A massa molecular relativa medida de um oligonucleotídeo é tipicamente de ~6.000, correspondendo a cerca de 20 bases. Sob condições em que os produtos amplificados de diferentes alelos contêm números diferentes de bases, a diferença de massa é ≥ 300, um valor significativo em relação à resolução da espectrometria de massa. De fato, é possível distinguir substituições de um único nucleotídeo em oligos de mesmo tamanho, ou ainda a metilação de um

sítio gPc. Para substituições de nucleotídeos, a diferença entre as massas varia de ± 9 para t↔a até ± 40 para c↔g.

As aplicações desta técnica incluem:

(1) **Medição de freqüências alélicas em populações ou detecção de alelos em indivíduos pela identificação de polimorfismos de um único nucleotídeo (SNPs)** Em estudos de populações, amostras de diversos indivíduos nos grupos selecionados podem ser analisadas e freqüências genotípicas podem ser medidas com uma precisão de cerca de 3%. Vários SNPs podem ser determinados a partir de um único espectro. Esses estudos têm impacto em muitas áreas, incluindo antropologia, agricultura e análises forenses, mas a área médica ainda é a principal beneficiária. Por exemplo, comparações controladas entre populações saudáveis e populações predispostas a uma doença podem levar à identificação de fatores genéticos de importância clínica.

(2) **Caracterização de genótipos individuais** Um grupo de 100.000 SNPs apresenta cerca de três polimorfismos por gene, o suficiente para a caracterização quase completa da porção codificadora de proteínas do genoma de uma pessoa. A determinação do perfil de SNPs de uma pessoa é possível utilizando-se um espectrômetro de massa por apenas um dia. As aplicações clínicas incluem (a) diagnósticos baseados em diferenças sistemáticas entre indivíduos saudáveis e doentes, estabelecidos previamente a partir de estudos de populações-controle e (b) farmacogenômica, para diferenciar pacientes que poderão se beneficiar do tratamento com um novo fármaco daqueles que não terão benefícios ou até mesmo estarão em risco de efeitos colaterais graves.

(3) **Medição de haplótipos individuais** Haplótipos são combinações locais de polimorfismos genéticos que tendem a ser herdados em conjunto (ver Quadro Distribuição de haplótipos). Os haplótipos simplificam as buscas por correlações entre fenótipo e genótipo, pois reduzem o número de variáveis necessárias para a caracterização do genótipo. Os métodos de espectrometria de massa baseados na amplificação de regiões próximas a SNPs, em amostras contendo uma única molécula de DNA, constituem um método rápido e preciso para a determinação em larga escala de haplótipos individuais.

Distribuição de haplótipos

Nossos genomas individuais são caracterizados pela distribuição de marcadores genéticos. Polimorfismos de um único nucleotídeo são características convenientes de serem observadas e estudadas em uma população ou entre populações. Apesar de a densidade média de SNPs em nossos genomas ser de aproximadamente 1 SNP/5 kbp, muitas regiões de 100 kpb apresentam um número menor de combinações possíveis de SNPs (tipicamente entre 2 e 4), sugerindo que a recombinação nessas regiões é um evento raro. Estes segmentos, que permanecem intactos, são separados por intervalos, nos quais a recombinação é mais freqüente.

Umas poucas combinações de SNPs definem o **haplótipo** de um indivíduo. O projeto internacional HapMap coleta e administra dados sobre a distribuição de haplótipos de diversas populações humanas.

Os haplótipos são difíceis de medir, pois é essencial determinar quais SNPs estão presentes em uma *mesma* cadeia de DNA. Claramente, estudos de amostras misturadas obtidas de diversos indivíduos podem determinar →

Distribuição de haplótipos (*continuação*)

freqüências de SNPs específicas, mas não a sua correlação com haplótipos. Mesmo uma amostra contendo os dois cromossomos equivalentes de uma única célula diplóide pode misturar as contribuições das duas cópias da região. No entanto, os estudos de espectrometria de massa a partir de uma única cópia de uma molécula de DNA amplificada, gerada por diluição, podem identificar a *combinação* de SNPs que aparecem juntas no mesmo cromossomo, permitindo uma caracterização sem ambigüidade do haplótipo.

(4) **Medição dos níveis de expressão gênica em escala absoluta com precisão de aproximadamente 3%** Isso é obtido misturando-se o RNA extraído de uma amostra com quantidades conhecidas de oligorribonucleotídeos relacionados e medindo-se as quantidades relativas de sinais dos oligos calibradores e dos naturais.

(5) **Diagnóstico pré-natal não-invasivo baseado em pequenas quantidades de DNA fetal presente no sangue materno** Por causa da presença de 90 a 95% de DNA materno na amostra, apenas as contribuições paternas para o feto podem ser identificadas. Entretanto, esta técnica é suficientemente sensível para detectar o gene *SRY*, demonstrando que o feto é do sexo masculino, ou outros alelos paternos que podem ser úteis no diagnóstico de anomalias genéticas. Deve ser enfatizado que o uso apenas de *amostras de sangue da mãe* evita os riscos significativos do procedimento invasivo de coleta de líquido amniótico.

(6) **Seqüenciamento de genomas** A espectroscopia de massa tem o potencial de competir em acurácia e desempenho com os métodos baseados em gel para a determinação em larga escala de seqüências de DNA.

Estudo de Caso 6.2: Aplicação combinada de genômica, proteômica e métodos estruturais em estudos de resistência a antibióticos em tuberculose

A tuberculose é uma doença infecciosa causada pelo *Mycobacterium tuberculosis*. Apesar do desenvolvimento de vacinas e drogas, ela ainda é uma doença letal. A tuberculose e a AIDS são as causas mais comuns de morte por doença infecciosa, com cerca de três milhões de vítimas por ano. A Organização Mundial da Saúde estima que ocorram oito milhões de novos casos anualmente. A infecção com HIV aumenta a mortalidade por tuberculose pela diminuição da resistência do paciente.

Nossa primeira linha de defesas contra a maior parte das infecções bacterianas inclui os macrófagos, células do sistema imune que engolem bactérias e as atacam com uma variedade de agentes químicos e bioquímicos. *M. tuberculosis* está extraordinariamente adaptado para viver *dentro* do macrófago. Parte da sua adaptação é estrutural: células de *M. tuberculosis* e outras bactérias relacionadas possuem um revestimento céreo, uma parede celular rica em ácidos graxos. A baixa permeabilidade desse revestimento protege a bactéria do ambiente inóspito no interior do macrófago, incluindo o baixo pH e o estresse oxidante. A bactéria também apresenta alterações substanciais no seu padrão de expressão gênica, adaptando o seu estado fisiológico às condições ambientais.

Após décadas de declínio em decorrência do desenvolvimento de drogas efetivas, a incidência de tuberculose começou a aumentar na década de 1980. Uma das razões para esta ressurgência foi o aparecimento de cepas resistentes.

O principal fármaco utilizado na prevenção e no tratamento da tuberculose é a isoniazida (hidrazida do ácido isonicotínico). A isoniazida combate o *M. tuberculosis* ao interferir na síntese de sua parede celular, sem a qual a bactéria não sobrevive. Os alvos da isoniazida incluem uma enoil-proteína carreadora de acila redutase NADH-dependente (InhA) e uma β-ceto-acil-proteína carreadora de acila sintase (KasA). Essas enzimas participam da síntese de ácidos micólicos, os principais componentes da parede celular.

A isoniazida precisa ser convertida na sua forma ativa após a absorção pela célula bacteriana. A enzima que catalisa esta conversão, KatG, é uma das suspeitas naturais de envolvimento com os mecanismos de resistência. Sua função natural é desintoxicar peróxidos.

Vários métodos foram utilizados para elucidar as adaptações responsáveis pela resistência à isoniazida:

(1) Alterações nos padrões de expressão gênica foram detectadas por microarranjos.

(2) Genes com alterações na expressão foram seqüenciados em cepas suscetíveis e resistentes, e as mutações observadas.

(3) A estrutura cristalina da isoniazida ligada à InhA foi determinada.

(1) Alterações nos padrões de expressão gênica

Wilson e colegas* examinaram cepas suscetíveis e resistentes de *M. tuberculosis* após 8 horas de exposição à isoniazida. (A Gravura IX mostra os resultados após 4 horas de exposição.) O arranjo incluiu quase todas as ORFs identificadas no genoma de *M. tuberculosis*. Apesar de estudos bioquímicos já terem implicado algumas proteínas na resistência, uma triagem completa foi realizada para identificar tantos alvos potenciais de fármacos quanto possível.

A exposição à isoniazida aumentou significativamente a transcrição de duas classes de genes. Um conjunto está envolvido com a síntese da parede celular, incluindo um grupo similar a um óperon codificador dos componentes do complexo de síntese de ácidos graxos (FAS-II). Genes adicionais, incluindo uma subunidade da alquil-hidroperóxido redutase C (AhpC), que controla o estresse oxidante, também apresentaram expressão aumentada. A lógica do experimento é que as células tratadas estão reconhecendo os efeitos do fármaco e mecanismos de regeneração estão agindo para tentar compensar a redução de atividade causada pelo fármaco, aumentando, assim, a expressão de determinados genes.

O genoma de *M. tuberculosis* tem cerca de 4,4 Mpb de extensão e contém aproximadamente 4.000 genes.

(2) Mutações que conferem resistência à isoniazida

Com base nas alterações dos perfis de expressão, Ramaswamy e colaboradores (2003) seqüenciaram um total de 2,6 Mpb de 124 isolados de *M. tuberculosis*.**

* Wilson, M., DeRisi, J., Kristensen H. H., Imboden, P., Rane, S., Brown, P. O. & Schoolnik, G. K. (1999), Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization, *Proc. Natl. Acad. Sci. USA*, **96**,12833-12838.

** Ramaswamy, S. V., Reich, R., Dou, S. J., Jasperse, L., Pan, X., Wanger, A., Quitugua, T. & Graviss, E. A. (2003), Single nucleotide polymorphism in genes associated with isoniazid. *Antimicrob. Agents Chemother*, **47**,1241-1250.

Estudo de Caso 6.2 (*continuação*)

Estes estudos incluíram as mutações na KatG que impedem a ativação da isoniazida e mutações na InhA que escapam à inibição mediada pela forma ativa do fármaco.

Note que, como o estresse oxidante é parte das defesas naturais do hospedeiro à infecção, um simples nocaute da KatG seria uma estratégia perigosa para a bactéria. De maneira ideal, para se tornar resistente, a bactéria precisaria reduzir a atividade da enzima na ativação da isoniazida, mas manter alguma atividade contra pequenos peróxidos. Dessa forma, ela reduziria a suscetibilidade ao fármaco e, ao mesmo tempo, manteria a sua adaptabilidade ao ambiente no interior do macrófago. Este balanço é obtido, de forma precisa, pela mutação mais comum da KatG observada em cepas resistentes, a S315T.

A mutação mais comum na InhA é a S94A. A eficácia inibitória da isoniazida ativada é reduzida nesta proteína modificada.

(3) Cristalografia

Rozwarski e colaboradores (1998) resolveram a estrutura do complexo entre a forma ativada da isoniazida e a InhA (Figura 6.5). O fármaco encontra-se covalentemente ligado ao anel nicotinamida do NAD e ambos estão ligados ao sítio ativo da InhA. A posição da cadeia lateral do resíduo S94 também é mostrada. No complexo inibitório, a proteína se liga ao aduto NAD-isoniazida ativada. A ligação entre o NAD e a isoniazida ativada pode ocorrer *somente* no sítio ativo da enzima (a isoniazida ativada e o NADH não reagem em solução).

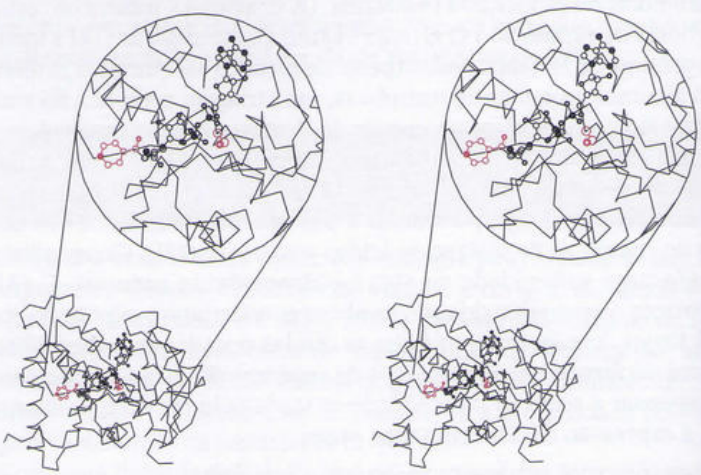


Figura 6.5 Estrutura da enoil-acp-redutase de ácidos graxos de cadeia longa (InhA) em complexo com seu inibidor [Código PDB: 1ZID]. O ligante é um aduto da forma ativada da isoniazida e NADH. Em vermelho estão a porção isoniazida do inibidor (à esquerda, no círculo destacado) e a cadeia lateral da Ser94 (ao centro, no círculo destacado). A mutação S94A contribui para a resistência à isoniazida. (Ver Rozwarski, D. A., Grant, G. A., Barton, D. H., Jacobs, W. R., Jr. & Sacchettini, J. C. (1998), Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. *Science*. **279**, 98-102.)

Como o mutante S94A se torna resistente? Na ausência do inibidor, a enzima pode se ligar primeiro ao substrato e depois à coenzima, ou vice-versa. Uma vez que o substrato ocupa o mesmo local do sítio ativo que o inibidor, apenas quando a coenzima se ligar primeiro o complexo de inibição poderá se formar. Duas vias são possíveis, a primeira origina exclusivamente o produto e a outra pode originar o complexo inibitório (E = enzima, S = substrato, C = coenzima, I = inibidor):

Se o substrato se liga primeiro:

(1a) $E + S \rightarrow ES + C \not\rightarrow ESC \rightarrow E + C + P$

(1b) $E + S \rightarrow ES + I \rightarrow$

Se a coenzima se liga primeiro:

(2a) $E + C \rightarrow EC + S \rightarrow ESC \rightarrow E + C + P$

(2b) $E + C \rightarrow EC + I \rightarrow ECI = \underset{\text{inibitório}}{\text{complexo}}$

Se o substrato se ligar primeiro (1a e 1b), o complexo inibitório não pode ser formado. Se a coenzima se ligar primeiro (2a e 2b), um complexo inibitório estável pode ser formado, eliminando a atividade desta enzima de maneira permanente.

A mutação S94A reduz a afinidade da enzima pelo NADH. Isso privilegia a via de ligação inicial ao substrato (1a e 1b), diminuindo a quantidade de complexos inibitórios produzidos (2b) e também aumentando a constante de dissociação do complexo inibitório.

É possível também que a mutação S94A e outras mutações reduzam a afinidade do aduto.

A pesquisa e o desenvolvimento de fármacos antituberculose são um desafio constante. Este exemplo mostra a eficácia da aplicação coordenada de várias técnicas diferentes.

Biologia de sistemas

A biologia de sistemas se concentra na integração e no controle da atividade de genes e de proteínas.

Proteínas são como animais sociais, e a vida depende de suas interações. Como proteínas individuais possuem funções especializadas, são necessários mecanismos de controle para a integração de suas atividades. A quantidade correta da proteína correta deve funcionar no local exato no momento exato. Falhas nos mecanismos de controle podem levar a doenças ou mesmo à morte.

Sob condições ambientais invariáveis, os sistemas bioquímicos de um organismo devem ser estáveis. Sob condições variáveis, o sistema deve ser robusto, acomodando tanto perturbações neutras quanto estressantes. Durante longos períodos de tempo, os processos devem ter suas taxas alteradas, ou mesmo serem ativados e desativados. Esta regulação inclui ajustes a curto prazo, por exemplo, nos estágios do ciclo celular, ou respostas a estímulos externos, como mudanças na composição ou concentração de nutrientes ou de oxigênio. Atividades reguladoras a longo prazo incluem o controle sobre os estágios de desenvolvimento durante toda a vida de um organismo.

O **metabolismo** é o fluxo de moléculas e energia através de vias de reações químicas. Muitas das reações metabólicas envolvem proteínas e ácidos

nucléicos, assim como pequenos componentes, como aminoácidos e açúcares. O conjunto completo de reações metabólicas forma padrões de tráfego complexos. Alguns desses padrões são vias lineares, como a síntese, em várias etapas, do triptofano a partir de corismato. Outros formam circuitos fechados, como a via do ácido tricarboxílico (ou ciclo de Krebs). Além disso, as vias metabólicas estão intimamente relacionadas. A estrutura da totalidade das vias metabólicas – sua conectividade ou topologia – e seus padrões de atividade podem ser analisados em termos de instrumentos matemáticos que lidam com grafos, fluxos e vazões.*

Para controlar os padrões de fluxo metabólico, **vias reguladoras** conectam proteínas e concentrações de metabólitos. A estrutura e a dinâmica das vias reguladoras são diferentes das observadas para as vias metabólicas. Correspondendo à sucessão de transformações enzimáticas no metabolismo, uma via reguladora é um agrupamento de cascatas de sinalização.

A biologia de sistemas descreve as interações metabólicas e reguladoras em termos de **redes de interações**.

Duas redes em paralelo: física e lógica.

Nas células, as duas redes de interações atuam em paralelo: (1) uma **rede física** de complexos proteína-proteína e proteína-ácido nucléico e (2) uma **rede lógica** de cascatas de controle. As vias metabólicas são compostas por ambas. Muitas, mas nem todas as vias metabólicas, são mediadas por interações físicas proteína-proteína e reguladas por interações lógicas.

Exemplos de interações puramente físicas incluem a formação dos centros de reação fotossintéticos, complexos de proteínas e co-fatores que convertem a luz em energia química, e a associação de fibras de colágeno no tecido conjuntivo. Exemplos de interações lógicas, *não* mediadas inteiramente pela interação física direta entre proteínas, incluem os ciclos de retroalimentação, nos quais o aumento na concentração de um produto de uma via metabólica inibe uma enzima que catalisa uma das etapas iniciais da via, ou a secreção de uma molécula pequena como sinal para outras células (modo de "disparo e esquecimento" (*"fire and forget"*). Nestes casos, a interação lógica é transmitida pela difusão da molécula pequena. Muitos outros exemplos aparecem no tema comum de regulação da expressão gênica. Um fator de transcrição, que se liga ao DNA, pode nunca interagir fisicamente com as proteínas cuja expressão ele controla.

Ver Quadro Comunicação célula-célula em microrganismos: percepção de *quorum*.

A mudança alostérica na hemoglobina é um exemplo de interações físicas e lógicas simultâneas: as subunidades da hemoglobina respondem a alterações nos níveis de oxigênio por meio de mudanças conformacionais que afetam a sua afinidade pelo oxigênio. Outro exemplo é a transmissão de um sinal da superfície de uma célula, através da membrana, até o seu interior, pela dimerização de um receptor. Este pode ser o gatilho inicial de uma cascata que, enfim, pode afetar a expressão gênica. Nem todas as etapas deste processo precisam envolver interações proteína-proteína; algumas podem ser mediadas pela difusão de pequenas moléculas, como o AMP cíclico.

Embora alguns complexos proteína-proteína e proteína-ácido nucléico participem tanto de redes físicas quanto de redes lógicas, as duas redes permanecem distintas em termos de seus fundamentos e de suas funções biológicas. É útil manter esta distinção em mente, *especialmente* quando são consideradas proteínas que atuam nas duas redes.

* N. de T. A relação entre o que entra e o que sai de um sistema.

Comunicação célula-célula em microrganismos: percepção de *quorum*

Mecanismos de controle que *não* envolvem interações proteína-proteína diretas medeiam a sinalização intercelular em microrganismos. A bactéria marinha *Vibrio fischeri* pode adotar estados fisiológicos alternativos, nos quais a bioluminescência pode estar ativa ou inativa (um "interruptor de luz", literalmente). O organismo pode viver livre na água do mar ou colonizar os órgãos geradores de luz de algumas espécies de peixes ou lulas. A bactéria é bioluminescente apenas quando cresce dentro desses animais.

A bactéria responde à densidade local de células, uma forma de comunicação denominada percepção de *quorum*. Em *V. fischeri*, a percepção de *quorum* é mediada pela secreção e detecção de uma pequena molécula sinalizadora, a N-(3-oxohexanoil)-homosserina lactona. Espécies relacionadas utilizam outras N-acil homosserina lactonas, abreviadas por AHL. A AHL pode se difundir livremente para fora da célula onde é sintetizada. Nos órgãos geradores de luz, a densidade pode chegar de 10^{10} a 10^{11} células/mL, e a concentração de AHL pode exceder o limiar de 5 a 10 nM para disparar o "interruptor" fisiológico.

Os genes bacterianos *LuxI* e *LuxR* são responsáveis por esta regulação. O produto do gene *LuxI* está envolvido na síntese de AHL. O produto do gene *LuxR* contém um domínio de ligação à membrana que detecta o sinal AHL e um domínio de ativação transcricional. *LuxR* ativa um óperon que inclui (1) genes para a síntese de luciferase (a enzima responsável pela bioluminescência) e (2) de *LuxI*, cuja expressão promove a síntese adicional de AHL, amplificando o sinal e acentuando a transição.

O hospedeiro também é sensível à bactéria: os órgãos geradores de luz de lulas crescidas em água salgada estéril não se desenvolvem de modo adequado. Aparentemente ocorre uma reação à luminescência e não à AHL. Para o animal, a luminescência contribui para a sua camuflagem, protegendo-o de predadores que habitam águas mais profundas, mesclando-se à luz do céu. O jogo de luz e sombras é uma forma natural de "maquilagem". (A bioluminescência freqüentemente surpreende pessoas, jantando em restaurantes especializados em frutos do mar, as quais podem concluir que seus jantares luminosos têm origem extraterrestre. No entanto, a maior parte das bactérias luminescentes é inofensiva, embora algumas cepas relacionadas à espécie *Vibrio*, como *Vibrio cholerae*, agente causador da cólera, sejam apenas fracamente luminescentes. A virulência de *V. cholerae* também está sob controle da percepção de *quorum* por um mecanismo relacionado.)

Redes e grafos

Em termos abstratos, as redes possuem a forma de grafos.

As rotas entre as cidades no mapa da Suécia na Figura 4.3 (página 197) são uma rede representada por um grafo, similar aos grafos que aparecem na biologia de sistemas. Cada cidade é um nó. As linhas grossas que as conectam indicam as estradas. Outros exemplos familiares a muitos leitores são os ma-

Ver Quadro A idéia de um grafo (página 335)

pas do metrô de Londres* e os mapas do sistema de metrô de outras cidades. Cada estação é um nó do grafo, e as arestas correspondem às linhas (trilhos) conectando as estações. O mapa moderno do metrô de Londres mostra a *topologia* da rede; ele não representa quantitativamente a geografia da área. Um mapa mais antigo, de 1925, manteve a precisão geográfica.** Isso era possível quando o sistema era mais simples do que o atual. Alguns dos mapas atuais do metrô de Paris são bastante fiéis geograficamente. *Considerados como redes, um mapa com fidelidade geográfica e um mapa simplificado, com a mesma topologia, correspondem ao mesmo grafo.*

A rede do metrô de Londres é totalmente conectada, no sentido de que existem conexões entre quaisquer duas estações. Muitas questões familiares aos usuários de metrô são compartilhadas com a análise de redes biológicas; por exemplo: quais são as vias que conectam as estações A e B? Considerando linhas diferentes como sub-redes, quão fácil é se transferir de uma sub-rede a outra, isto é, qual é a natureza dos padrões de conectividade? No caso de falha de uma ou mais conexões, a rede ainda se mantém robusta (ou seja, ela se mantém totalmente conectada)?

Sistemas biológicos precisam ser robustos, tanto para a sobrevivência de organismos sob estresse quanto para a plasticidade necessária à evolução. Em levedura, por exemplo, nocautes de um único gene, em mais de 80% das cerca de 6.200 ORFs, causam danos que não afetam a sobrevivência do organismo.

Em princípio, as redes se tornam robustas por meio de redundância. O mecanismo mais direto é uma simples **redundância de substituição**: se duas proteínas são capazes de realizar uma tarefa, se uma for eliminada, a outra realizará a função. Na analogia com o metrô de Londres, isso corresponderia a uma segunda linha passando pela mesma rota. Por exemplo, quando a linha Circle Line não estiver funcionando, os passageiros que estiverem viajando entre as estações Paddington e King's Cross podem utilizar a linha Hammersmith & City Line, que passa pelos mesmos trilhos.

Nas células, alguns genes possuem homólogos estreitamente relacionados, originados por duplicação gênica, e alguns desses homólogos contribuem para a redundância de substituição. Por exemplo, em estudos com modelos de camundongo para diabetes, descobriu-se que camundongos e ratos (mas não humanos) possuem dois genes similares, mas não alélicos, para a insulina. Contudo, a redundância de substituição exige não apenas a equivalência de funções, mas também de controle de expressão. No camundongo, o nocaute de qualquer um dos genes para insulina leva ao aumento compensatório da expressão do outro gene e um fenótipo normal. Padrões de expressão equivalentes são mais prováveis entre genes duplicados do que entre genes não relacionados. Por exemplo, a bactéria *E. coli* contém duas proteínas aldolase frutose-1,6-bifosfato. Uma é expressa apenas na presença de nutrientes especiais e não é essencial para o crescimento sob condições normais. A outra, porém, é essencial. Neste caso, a redundância funcional *não* provê mais robustez. Estas duas enzimas são prováveis homólogas, com parentesco distante, e não o produto de uma duplicação gênica recente. Uma delas é membro de uma família das aldolases frutose-1,6-bifosfato típicas de bactérias e eucariotos. A outra é membro de uma outra família que ocorre em arqueobactérias. *E.coli* é uma exceção pelo fato de conter as duas proteínas.

* `http://www.transportforlondon.gov.uk/tfl/tube_map.shtml` ou `http://www.afm.org/~alplatt/tube.html`. Os Exercícios 6.11 e 6.12, o Problema 6.5 e o *Weblema* 6.4 também utilizam este mapa.

** `http://www.ltmuseum.co.uk/collections/collections/posters.shtml`

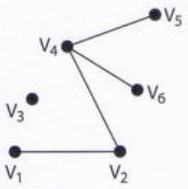
Um mecanismo alternativo para aumentar a robustez de uma rede é a **redundância distribuída**: o mesmo efeito é obtido por meio de rotas diferentes. Em bactérias *E. coli* normais, cerca de dois terços do NADPH produzido no metabolismo originam-se através da rota de salvamento das pentoses fosfato, a qual requer a enzima glicose-6-fosfato desidrogenase. O nocaute do gene para essa enzima leva a desvios metabólicos, nos quais o excesso de NADH produzido pelo ciclo dos ácidos tricarboxílicos é convertido em NADPH por uma reação de transidrogenase. A taxa de crescimento das cepas nocaute é comparável com a das cepas normais.

A idéia de um grafo

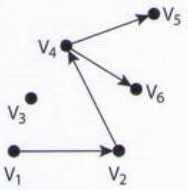
- Matematicamente, um grafo consiste em um conjunto de vértices V e um conjunto de arestas E.
- Cada aresta é especificada por um par de vértices.
- Em um **grafo orientado**, as arestas são pares **ordenados** de vértices.
- Em um **grafo rotulado**, existe um valor associado a cada aresta (um grafo orientado é um caso especial de grafo rotulado: considere a ponta das setas como sendo os rótulos).

Um grafo não-orientado e não-rotulado especifica a conectividade de uma rede, mas não as distâncias entre os vértices (especifica a topologia, mas não a geometria, como nos mapas modernos do metrô de Londres). Rótulos nas arestas podem indicar distâncias. Por exemplo, algumas árvores filogenéticas indicam apenas a topologia da ancestralidade. Outras indicam quantitativamente o grau de divergência entre as espécies. Árvores filogenéticas são em geral desenhadas com os comprimentos dos ramos indicando o tempo decorrido desde o último ancestral comum. Esta é uma forma pitoresca de rotular as arestas de um grafo.

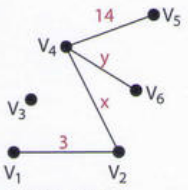
Alguns grafos não correspondem a estruturas físicas e, em todo o caso, os rótulos das arestas não precisam refletir a geometria no seu sentido comum. Por exemplo, as conexões em uma rede de vias metabólicas poderiam ser rotuladas para indicar os padrões de fluxo de metabólitos.



grafo



grafo orientado



grafo rotulado

Exemplos de grafos

- Grupos de pessoas que se conhecem
- Sistemas de distribuição de eletricidade
- Árvores filogenéticas
- Vias metabólicas
- Padrões de ligações químicas em moléculas
- Padrões de citações na literatura científica
- A rede mundial de computadores (*World Wide Web*)

Uma propriedade fundamental de uma rede é a sua **conectividade**.

Se V_A e V_Z são vértices em um grafo, o **caminho** de V_A até V_Z é uma série de vértices: V_A, V_B, V_C, ... V_Z, de forma que uma aresta no grafo conecta cada par sucessivo de vértices. O número de vértices em uma cadeia é chamado de **comprimento** do caminho. Um **ciclo** é um caminho de comprimento >2 em um grafo não-orientado, no qual o ponto inicial e final é o mesmo, mas nenhuma conexão intermediária se repete.

Um grafo que contenha um caminho entre dois vértices quaisquer é chamado de grafo **conexo**. Alternativamente, um grafo pode se dividir em vários componentes conexos denominados ***cliques***. O grafo na margem da página 335 contém dois *cliques*, um contendo cinco vértices e outro contendo apenas um vértice. (De forma extrema, um grafo pode conter muitos vértices e nenhuma aresta.) Geralmente é útil determinar o *caminho mais curto* entre dois nós quaisquer e caracterizar uma rede pela distribuição de comprimentos de caminhos. A declaração bastante conhecida de que quaisquer duas pessoas estão conectadas por "seis graus de separação" significa que, se as pessoas do mundo são vértices em um grafo e o grafo contém uma aresta sempre que duas pessoas se encontram, então o grafo é totalmente conexo e existe um caminho entre quaisquer dois vértices cujo comprimento é ≤ 6.

Uma **árvore** é uma forma especial de grafo. Uma árvore é um grafo conexo contendo apenas um caminho entre cada par de vértices. Uma hierarquia é uma árvore: exemplos incluem as cadeias militares de comando e a taxonomia de Lineu. Note que algumas árvores genealógicas não são árvores no sentido matemático: exemplos sobram nas famílias reais européias. Uma árvore não pode conter um ciclo: se contivesse, haveria dois caminhos a partir do ponto inicial (= ao ponto final) para cada ponto intermediário. No grafo da margem da página 335, o subgrafo formado pelos vértices V_1, V_2, V_4, V_5 e V_6 é uma árvore. A adição de uma aresta entre V_1 e V_5 criaria um caminho alternativo de V_1 a V_5 e o ciclo $V_1 \rightarrow V_2 \rightarrow V_4 \rightarrow V_5 \rightarrow V_1$, e o grafo deixaria de ser uma árvore.

A **densidade de conexões**, ou seja, o número médio de arestas por vértice caracteriza a estrutura do grafo. Um grafo de N vértices totalmente conexo possui $N - 1$ conexões por vértice; um grafo sem arestas possui um número de conexões igual a zero. O sistema nervoso dos animais superiores atingiu a sua capacidade não apenas pelo grande número de neurônios, mas também pela sua alta conectividade.

Em alguns sistemas existe um limite no número de conexões: para muitas sociedades humanas, em um grafo onde as pessoas são os vértices e as arestas unem as pessoas casadas, cada nó possui conectividade 0 ou 1. Em hidrocarbonetos, os grafos nos quais os átomos de carbono e de hidrogênio são os vértices e as arestas unem os átomos ligados covalentemente, cada nó possui um número de conexões ≤ 4. Em outras redes, a conectividade segue regularidades observáveis (ver Quadro Redes do tipo "mundo pequeno"). Por exemplo, a rede mundial de computadores (*World Wide Web*) pode ser considerada um grafo orientado. Documentos individuais são os nós, e as hiperconexões são as arestas. Observou-se que a distribuição de conexões de entrada e saída seguem as leis de potência: $P(k)$ = probabilidade de que k arestas sejam proporcionais a k^{-q}, onde $q = 2{,}1$ para conexões de entrada e $q = 2{,}45$ para conexões de saída.

Redes do tipo "mundo pequeno"

Muitas redes, incluindo redes biológicas, a rede mundial de computadores e redes de distribuição de eletricidade, apresentam as características de alto agrupamento e pequenos comprimentos de caminhos. Elas incluem relativamente poucos nós com um alto número de conexões chamadas de "roteadores" (*hubs*), e muitos nós com poucas conexões. Esses nós se combinam para gerar caminhos curtos entre todos os nós. Por essa razão, tais redes são chamadas redes do tipo "mundo pequeno". Essas redes tendem a ser bastante robustas – permanecendo conexas mesmo após uma falha aleatória de nós. A falha de um dos roteadores seria desastrosa, mas é improvável em função da existência de poucos roteadores.

Muitas redes, sobretudo a rede mundial de computadores, estão adicionando nós continuamente. A distribuição da conectividade tende a permanecer razoavelmente constante conforme a rede cresce. Essas redes são chamadas de **redes "sem escala"**.

A densidade de conexões é muito importante na definição das propriedades de uma rede. Por exemplo, as interações que espalham doenças entre os homens e entre os animais formam uma rede. Se uma doença será capaz de causar uma epidemia depende não apenas da facilidade de transmissão em qualquer interação particular, mas também da densidade de conexões. Conforme a densidade de conexões – a taxa de interações – aumenta, o sistema pode exibir uma mudança *qualitativa* de comportamento, análoga à mudança de fase em físico-química, de uma situação em que a doença permanece sob controle para uma epidemia que se espalha por toda a população. O procedimento clássico de quarentena – isolamento de pessoas por quarenta dias – funciona ao diminuir o grau de conectividade da rede de transmissão da doença. Note que um portador assintomático – o tifo é um caso clássico, com a "Maria tifóide"* – serve como um roteador para a rede de transmissão da doença.

Duas epidemias históricas associadas a guerras demonstram a distinção entre topologia e geometria na conectividade de redes. (1) Nos primeiros anos da guerra do Peloponeso, Atenas sofreu uma epidemia grave. (Apesar da descrição detalhada dos sintomas feita por Tucídides, a doença não foi caracterizada, mas se tratava provavelmente da peste bubônica.) Um fator que contribuiu para a sua transmissão foi a aglomeração de pessoas na cidade, abandonando a maior vulnerabilidade do meio rural. (2) Após a Primeira Guerra Mundial, uma epidemia de influenza matou cerca de 20 milhões de pessoas, número de mortes superior ao registrado na própria guerra. As viagens de longa distância dos soldados que retornavam da guerra ajudaram na disseminação da doença. Toda epidemia precisa de um agente infeccioso e de uma alta densidade de rotas de transmissão. O fator de controle é a densidade de *conexões* e não a densidade de pessoas.

Uma mudança de comportamento análoga à transição para uma epidemia é observada na fissão nuclear. Em uma amostra de urânio^{235}, o decaimento do nú-

* Mary Mallon (1869-1938) apresentava a seguinte combinação desafortunada de características: (1) estava infectada com tifo, (2) não apresentava sintomas e (3) trabalhava como cozinheira para muitas famílias.

cleo produz nêutrons que podem iniciar a fissão de outros núcleos. Se a amostra for pequena, muitos nêutrons secundários são perdidos através da sua superfície e a amostra se mantém estável. Acima de uma massa crítica, uma quantidade suficiente de nêutrons é capturada no interior da amostra para criar uma reação em cadeia. Se os átomos são os vértices de um grafo e as arestas são as trajetórias dos nêutrons de um átomo ao outro, a mudança de comportamento pode ser vista como o efeito de aumentar a conectividade de uma rede. (O tema da peça popular recente de Michael Frayn, Copenhagen, envolve as tentativas, antes e durante a Segunda Guerra Mundial, de se estimar o tamanho da massa crítica para estabelecer se as armas nucleares seriam ou não uma realidade.)

Estrutura e dinâmica de redes

Um grafo não-rotulado e não-orientado fornece a estrutura *estática* da topologia de uma rede. Para redes de interações moleculares, esta pode ser uma descrição inadequada das muitas interações físicas.

Para algumas redes, como vias metabólicas ou padrões de tráfego em cidades, a *dinâmica* do sistema depende das capacidades de transmissão de cada uma das conexões entre os nós. Essas capacidades podem ser indicadas como rótulos nas arestas dos grafos. Isso permite a modelagem dos padrões de fluxo através da rede. Exemplos incluem o planejamento de rotas em viagens ou entregas. Note que a via mais curta pode não representar o fluxo ótimo. Motoristas de táxis são extremamente sensíveis – e, em alguns casos, extremamente volúveis – sobre as melhores rotas de tráfego.

Na biologia molecular, as rotas metabólicas e as cascatas de sinalização são redes que permitem análises por meio dos conceitos de vias e de fluxo. Mesmo o alinhamento ótimo de seqüências por programação dinâmica (ver Capítulo 4) envolve a determinação de um caminho ótimo através de um grafo de edição.

Apesar de haver um grande conhecimento acerca dos mecanismos de elementos individuais de vias de controle e de sinalização, a compreensão de sua integração é objeto de pesquisa atual. Por exemplo, a idéia de que células e organismos saudáveis se encontram em estados estáveis não é mais do que uma aproximação (e em muitos casos uma idealização). A descrição de um estado dinâmico real das redes metabólicas e reguladoras é um problema bastante delicado. Entender *como* as células atingem a estabilidade (mesmo sendo uma aproximação aparente) também é bastante complexo. É provável que uma grande redundância de processos de controle seja a base desta estabilidade. A regulação é o resultado de muitos mecanismos individuais de controle – um pequeno laço de retroalimentação aqui, uma cascata de múltiplas etapas ali. De alguma forma, as ações independentes de cada um dos sinais individuais se combinam para atingir um resultado global integrado. É como a operação da "mão invisível" que, segundo Adam Smith, coordena o comportamento individual na regulação das economias nacionais.

Vários tipos de estados dinâmicos de uma rede são possíveis (ver Quadro):

- Equilíbrio
- Estado estacionário
- Estados que variam periodicamente
- Desencadeamento de programas de desenvolvimento de um organismo
- Estados caóticos
- Descontrole ou divergência
- Inativação

Estados de uma rede de processos

- Em **equilíbrio,** um ou mais processos em um sentido e no sentido reverso ocorrem em taxas compensatórias, de forma que as quantidades de diferentes substâncias permanecem inalteradas:

$$A \rightleftarrows B$$

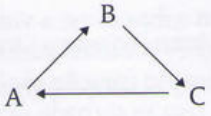
Equilíbrios químicos geralmente são capazes de auto-ajuste após mudanças nas condições ou nas concentrações de reagentes e produtos.

- Um **estado estacionário (*steady state*)** existirá se a taxa total de processos que produzem uma substância for igual à taxa total de processos que a consomem. Por exemplo, a conversão em duas etapas:

$$A \rightarrow B \rightarrow C$$

pode manter a quantidade de B constante se a taxa de produção de B (o processo $A \rightarrow B$) for a mesma do seu consumo (o processo $B \rightarrow C$). O efeito geral é a conversão de A em C.

Um processo cíclico poderia manter um estado estacionário em todos os seus componentes:



Um estado estacionário em um processo cíclico como este, em que todas as reações ocorrem em um mesmo sentido, é bastante diferente de um estado de equilíbrio. Entretanto, em alguns casos, a alteração das condições externas produz um desvio para outro estado estacionário adjacente e similar.

- **Estados que variam periodicamente** são encontrados na regulação do ciclo celular, no ritmo circadiano e nas mudanças sazonais, como os padrões anuais de acasalamento de animais e de florescência das plantas. Ciclos circadianos e sazonais têm suas origens nas progressões regulares de dias e anos, mas desenvolveram um certo grau de internalização.
- Muitas condições de equilíbrio e estado estacionário são **estáveis**, no sentido de que as concentrações da maioria dos metabólitos estão mudando lentamente, se de fato houver alguma mudança, e de o sistema ser capaz de responder bem a pequenas variações nas condições externas. O estado alternativo é um **estado caótico**, no qual pequenas variações nas condições podem causar grandes respostas. O tempo (estado atmosférico) é um sistema caótico: o meteorologista Lorenz certa vez perguntou "Será que o bater das asas de uma borboleta no Brasil causa um tornado no Texas?" Em um sistema cuidadosamente regulado, o caos é normalmente evitado, e é provável que a vida tenha evoluído para controlar as respostas aos tipos de flutuações que poderiam originar um estado caótico. Uma dinâmica caótica pode, em alguns casos, aproximar-se de estados estáveis – o que é chamado de **atratores estranhos (*strange atractors*)**. Compreender a estabilidade de sistemas dinâmicos sujeitos a estímulos ambientais variáveis é um assunto muito importante, mas que está além do escopo deste livro.

→

Estados de uma rede de processos (*continuação*)

- O **desencadeamento de programas de desenvolvimento** ocorre ao longo do curso da vida de uma célula ou organismo. Muitos eventos desenvolventes são relativamente independentes de condições externas e são controlados sobretudo pela regulação de padrões de expressão gênica.
- **Descontrole** ou **divergência**. A perda do controle sobre a proliferação celular leva ao crescimento desenfreado no câncer.
- A **inativação** é apenas uma parte do quadro geral. A apoptose é a morte programada de uma célula como parte do processo normal de desenvolvimento ou em resposta a danos que poderiam colocar em risco a integridade do organismo, como a quebra das cadeias de DNA. A perda dos mecanismos de apoptose – por exemplo, mutações na proteína p53 – é uma das principais causas de câncer.

Complexos e agregados de proteínas

A base do nosso conhecimento sobre como a vida dentro de uma célula é organizada é o conjunto de interações proteína-proteína e proteína-ácido nucléico. O desenvolvimento de métodos de geração rápida e em larga escala de dados para detectar essas interações tem se tornado um foco de interesse recente.

As interações que envolvem proteínas e ácidos nucléicos apresentam uma ampla gama de estruturas e funções.

- Dímeros simples ou oligômeros em que os monômeros parecem funcionar de maneira independente.
- Oligômeros com "interação cruzada" funcional, incluindo a dimerização de receptores induzida por ligantes e proteínas alostéricas, como a hemoglobina, fosfo-frutocinase e aspartato carbamoil-transferase.
- Grandes proteínas fibrosas como a actina ou a queratina.
- Agregados estruturais não-fibrosos, como os capsídeos virais.
- Grandes agregados com propriedades dinâmicas, como F1-ATPase, piruvato cinase, chaperonina GroEL-GroES e o proteossomo.
- Complexos proteína-ácido nucléico, incluindo ribossomos, nucleossomos, complexos de regulação de transcrição, complexos de recomposição (*splicing*) e reparo, bem como vírus. Em muitos casos, a ligação inicial é seguida pelo recrutamento de proteínas adicionais para formar complexos ainda maiores.
- Muitas proteínas, sejam monoméricas ou oligoméricas, funcionam ao interagir com outras proteínas. Exemplos de tais proteínas incluem todas as enzimas com substratos protéicos e muitos anticorpos, inibidores e proteínas reguladoras.
- Interações envolvendo proteínas são freqüentemente associadas a doenças, pois formas enoveladas erroneamente, ou proteínas mutantes, têm a tendência de formar agregados (ver Quadro Doenças associadas com agregados protéicos).

Doenças associadas com agregados protéicos

Doença	Proteína agregadora	Comentários
Anemia falciforme	Desoxi-hemoglobina S	Mutações criam regiões hidrofóbicas na superfície
Amiloidoses clássicas	Cadeias leves das imunoglobulinas, transtiretina e muitas outras	Depósitos fibrilares extracelulares
Enfisema associado ao alelo Z da antitripsina	Mutante da antitripsina α_1	A desestabilização da estrutura facilita a agregação
Huntington	Huntingtina alterada	Uma das doenças graves associadas com repetições de poliglutamina
Parkinson	α-sinucleína	Encontrada nos corpos de Lewy
Alzheimer	Aβ, τ	Aβ = fragmento de 40-42 resíduos
Encefalopatias espongiformes	Proteínas priônicas	Infecciosa, apesar de não conter ácidos nucléicos

Propriedades de complexos proteína-proteína

Estequiometria – qual é a composição do complexo?

Proteínas oligoméricas estáveis podem conter muitas cópias de uma única proteína ou combinar cópias de proteínas diferentes. Dentre os agregados de um único tipo de proteína, os complexos contendo um número *ímpar* de moléculas são menos comuns do que os complexos com um número par. Oligômeros (complexos contendo poucas cópias da mesma proteína – dímeros, trímeros, etc.) geralmente apresentam simetria. Por exemplo, a insulina é um hexâmero, e cada subunidade do hexâmero correlaciona-se de forma simétrica com as demais através de um eixo de ordem 3 perpendicular a outro eixo de ordem 2.

Algumas proteínas procarióticas que contêm subunidades idênticas são homólogas a proteínas eucarióticas que contêm subunidades relacionadas, mas não idênticas, originadas por duplicação gênica e divergência. O proteossomo é um exemplo. Alguns vírus alcançam a diversidade *sem* duplicação, ao combinar proteínas de mesma seqüência, mas com conformações diferentes.

Os complexos de proteínas variam bastante em número e variedade de moléculas que contêm. Alguns complexos contêm apenas poucas proteínas, mas outros são muito grandes: por exemplo, a piruvato desidrogenase contém centenas de subunidades, e alguns capsídeos virais são formados por milhares de subunidades.

A maioria dos agregados muito grandes tem importância clínica, incluindo a encefalopatia espongiforme bovina (BSE – Bovine Spongiform Encephalopathy, ou doença da vaca louca), o mal de Alzheimer e a doença de Huntington. As amiloidoses são doenças caracterizadas por depósitos fibrilares extracelulares, geralmente com uma estrutura comum de folha β cruzada. Esses depósitos surgem por uma variedade de causas, inclusive mutações desestabilizadoras, superprodução de uma proteína e eliminação inadequada em casos de falha renal. Proteínas desenoveladas são mais propensas a formar agregados, e proteínas mutadas são mais propensas a terem um enovelamento errôneo. Grandes concentrações

Ver Quadro Doenças associadas com agregados protéicos

locais, como nos casos de mielomas, que produzem altas quantidades de cadeias leves de imunoglobulinas, também favorecem a formação de agregados.

Afinidade – quão estável é o complexo?

Um indicador comum da afinidade de um complexo é a **constante de dissociação**, K_D, a constante de equilíbrio para o *inverso* da reação de ligação:

$$\text{Proteína} \cdot \text{Ligante} = \text{Proteína} + \text{Ligante} \qquad K_D = \frac{[p][l]}{[pl]}$$

A constante de Michaelis de uma enzima é a constante de dissociação do complexo enzima-substrato.

[P], [L] e [PL] denotam os valores numéricos das concentrações de proteína, ligante e complexo proteína-ligante, respectivamente, expressos em mol ℓ^{-1}. Quanto menor a K_D, mais forte a ligação. O valor de K_D corresponde à concentração de ligante livre na qual metade das proteínas está ligada ao ligante e metade se encontra livre em solução: [P] = [PL].

A constante K_D está relacionada com a mudança na energia livre de Gibbs para uma reação de dissociação pela seguinte relação:

$$PL = P + L \qquad \Delta G^{\ominus} = \Delta H^{\ominus} - T\Delta S^{\ominus} = -RT \ln K_D$$

As constantes de dissociação de complexos proteína-ligante apresentam uma ampla variação:

Contexto biológico	Ligante	Valor típico de K_D
Ativador alostérico	Íon monovalente	10^{-4} a 10^{-2}
Ligação de coenzima	NAD, por exemplo	10^{-7} a 10^{-4}
Complexos antígeno-anticorpo	Vários	10^{-4} a 10^{-11}
Inibidor de trombina	Hirudina	5×10^{-14}
Inibidor de tripsina	Inibidor da tripsina pancreática bovina	10^{-14}
Streptavidina	Biotina	10^{-15}

Estudos estruturais elucidaram diversas características importantes das interações entre proteínas solúveis que contribuem para a afinidade:

- **O que mantém as proteínas ligadas?** O efeito hidrofóbico, ligações de hidrogênio e pontes salinas.
- **As proteínas mudam de conformação durante a formação dos complexos?** Em alguns casos, sim. Nestes casos a energia de interação tem de "pagar" pela mudança conformacional, e a interface tende a ser maior.
- **O que determina a especificidade?** A complementaridade das superfícies de contato – em termos de forma, potencial de formação de ligações de hidrogênio e distribuição de cargas (ver Quadro Características de interfaces proteína-proteína). A predição de complexos protéicos a partir da estrutura das proteínas que os compõem é um **problema de docagem**. Uma solução confiável para esse problema, juntamente com os progressos na genômica estrutural, permitirão a triagem *in silico* de proteomas em busca de proteínas que participem de interações.

Cinética de formação e dissociação, tempo de meia-vida

A constante de dissociação de um complexo indica a fração de tempo em que os seus componentes passam no estado ligado e a fração de tempo em que eles não

estão ligados. Porém, o **tempo de meia-vida** do estado ligado pode variar sem afetar K_D. Definindo taxas individuais para associação e dissociação, k_{on} e k_{off}, a constante de dissociação é igual à razão entre elas:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} [PL] \quad K_D = k_{off}/k_{on}$$

Uma meia-vida curta, correspondendo a valores elevados de k_{off} e k_{on}, ou uma meia-vida longa, correspondente a valores pequenos de k_{off} e de k_{on}, podem originar o mesmo K_D. Os valores do tempo de meia-vida são importantes: se você quiser purificar um complexo, é importante que a sua meia-vida seja maior do que a duração do procedimento de isolamento! Ao contrário, no caso de um complexo proteína-proteína que medeie a transmissão de um sinal, uma meia-vida curta fornece um "mecanismo de reinicialização" natural, evitando que o sinal permaneça "ligado" por um período de tempo muito longo.

Características de interfaces proteína-proteína

- **Proteção de superfície da proteína.** A área acessível da superfície (ASA – Accessible Surface Area) de uma proteína é calculada deslizando-se uma esfera sonda, do tamanho de uma molécula de água (raio igual a 1,4 Å), sobre a sua superfície. A superfície protegida pela formação de um complexo é a diferença entre a ASA do complexo e a soma das ASAs de cada um dos componentes separadamente.

  Uma interface proteína-proteína típica pode envolver 22 resíduos e 90 átomos, dos quais 20% seriam átomos da cadeia principal e, ocasionalmente, uma molécula de água. Um histograma da área da superfície protegida em complexos binários de proteínas mostra um pico centrado em 1.600 $Å^2$.

  A superfície protegida mínima para a estabilidade de um complexo proteína-proteína é de cerca de 1.000 $Å^2$. Complexos com superfícies protegidas com mais de 2.000 $Å^2$ tendem a envolver mudanças conformacionais das subunidades durante a sua formação.

  Cada $Å^2$ de superfície hidrofóbica protegida contribui com cerca de 105 J mol^{-1} para a energia livre de estabilização.
- **A composição da interface.** O caráter químico das interfaces de proteínas é intermediário entre o das superfícies e o dos interiores de proteínas globulares monoméricas. As interfaces são ricas em átomos polares neutros à custa de átomos carregados. As composições de aminoácidos das interfaces são ricas em resíduos aromáticos – His, Phe, Tir e Trp – em relação ao restante da superfície exposta. Existe ainda uma menor quantidade de cadeias laterais alifáticas – Leu, Ile, Val, Met – e de resíduos de Arg (mas surpreendentemente não de Lis).
- **A complementaridade** das interfaces é responsável pela especificidade. A complementaridade envolve tanto um bom empacotamento das superfícies oclusas quanto uma justaposição apropriada de átomos envolvidos em ligações de hidrogênio e de átomos carregados. Existe tipicamente uma ligação de hidrogênio para cada 170 $Å^2$ de área da interface. Moléculas de água isoladas estão presentes em muitas interfaces. Existe, em geral, uma molécula de água para cada 100 $Å^2$ de interface.

A taxa de associação (*"on rate"*) é limitada pelas taxas de difusão. Sob condições normais $k_{on} \leq 10^{-9}$ $M^{-1}s^{-1}$. Se uma mudança conformacional for necessária para a ligação, o valor de k_{on} pode ser significativamente menor. Tipicamente, as taxas k_{on} estão entre 10^{-6} e 10^{-7} $M^{-1}s^{-1}$, com meia-vida de aproximadamente 1 s.

Como os complexos são organizados em três dimensões?

Quando duas proteínas formam um complexo, cada uma deixa uma "impressão" na superfície da outra, definindo a porção da superfície envolvida na interação. Se duas proteínas interagem por meio da *mesma* superfície em cada uma delas, o complexo é denominado **fechado**. Se duas proteínas interagem através de superfícies *diferentes*, o complexo é dito **aberto**. A importância desta distinção é que complexos fechados não permitem a ligação de proteínas adicionais com a mesma interação. Um complexo aberto, no qual a superfície de interação em potencial não está oclusa, pode crescer pela adição de mais subunidades. Dessa forma, os complexos abertos, mas não os fechados, são compatíveis com a formação de agregados pela replicação da interação.

As proteínas mudam de conformação durante a formação do complexo?

Alguns complexos de proteínas se formam pela união de subunidades rígidas. As subunidades nesses complexos possuem a mesma estrutura quando unidas e quando separadas. Outros complexos de proteínas envolvem mudanças estruturais durante a sua formação. Tais complexos incluem aqueles em que as subunidades independentes não são estáveis separadamente.

Redes de interações protéicas

As unidades que formam redes de interações são:

- Para redes físicas, um complexo proteína-proteína ou proteína-ácido nucléico.
- Para redes lógicas, uma conexão dinâmica na qual a atividade de um processo é afetada por uma alteração nas condições externas ou pela atividade de outro processo.

Muitos experimentos revelam apenas interações par a par. Os desafios são integrar as interações par a par em uma rede e, então, estudar a estrutura e a dinâmica do sistema.

Muitas técnicas detectam interações físicas diretamente. Essas técnicas incluem as seguintes:

- A **determinação de estruturas por raios X e por NMR** pode não apenas identificar os componentes do complexo, mas também revelar como as subunidades interagem e se ocorrem mudanças conformacionais durante a formação do complexo.
- **Sistemas de triagem de duplo-híbrido**. Ativadores da transcrição, como Gal4, contêm um domínio de ligação ao DNA e um domínio de ativação. Suponha que esses dois domínios sejam separados e que uma proteína-teste seja fusionada ao domínio de ligação ao DNA e uma segunda proteína-teste seja fusionada ao domínio de ativação. Uma proteína "repórter" será expressa apenas se os componentes do ativador forem unidos pela formação de um

complexo entre as duas proteínas-teste. Métodos de geração rápida e em larga escala de dados permitiram a triagem em paralelo de uma proteína "isca" por interações com um grande número de potenciais proteínas "predadoras" (ver Quadro Redes de interações entre proteínas detectadas pelos sistemas de triagem de duplo-híbrido).

Redes de interações entre proteínas detectadas pelos sistemas de triagem de duplo-híbrido*

	H. pylori	*S. cerevisiae*	*C. elegans*	*D. melanogaster*
Tamanho total do proteoma	1.576	5.585	33.469	13.843
Proteínas testadas	732	987 / 790	1.415	4.685
Interações detectadas	1.465	936 / 800	2.131	4.876

Os dois conjuntos de números para levedura são o resultado de estudos independentes.

* Adaptado de Aloy, P. & Russell, R.B. (2004), Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, **22**,1317-1321.

- A **ligação química cruzada** liga complexos de forma que eles possam ser isolados. A subseqüente digestão proteolítica e a espectrometria de massa permitem a identificação dos componentes.
- **Co-imunoprecipitação.** Um anticorpo desenvolvido para a proteína "isca" se ligará à "isca" juntamente com quaisquer outras proteínas "predadoras" que interajam com ela. As proteínas ligadas podem ser purificadas e analisadas, por exemplo, por Western *blotting* ou espectrometria de massa.
- A **imunoprecipitação de cromatina** identifica seqüências de DNA que se ligam a proteínas. O tratamento com formaldeído fixa (com ligações cruzadas) os complexos proteína-DNA que existem na célula. Então, o isolamento da cromatina e a clivagem do DNA em pequenos fragmentos permitem a separação de proteínas pela ligação a anticorpos específicos, mantendo a seqüência de DNA ligada ao complexo. O rompimento das ligações cruzadas, seguido pelo seqüenciamento do DNA, identifica a seqüência de DNA específica à qual cada proteína se liga.
- **Exibição por fagos.** Genes para um grande número de proteínas são fusionados individualmente ao gene para uma proteína de revestimento de um fago, criando uma população de fagos, na qual cada fago carrega cópias de uma das proteínas extras expostas em sua superfície. A purificação por afinidade empregando proteínas "isca" imobilizadas seleciona os fagos que expõem potenciais proteínas "predadoras". A extração de DNA dos fagos interagentes revela as seqüências de aminoácidos dessas proteínas.
- A **ressonância plasmônica de superfície** analisa a reflexão da luz em uma superfície de ouro à qual a proteína foi fixada. O sinal muda caso um ligante se ligue à proteína imobilizada. (O método detecta mudanças locais no índice de refração do meio adjacente à superfície de ouro. Esta variação está relacionada com a massa da proteína imobilizada.)

- **Transferência de energia ressonante fluorescente (FRET).** Se duas proteínas forem marcadas com cromóforos diferentes, a transferência da energia de excitação pode ser observada por meio de distâncias que podem chegar a 60 Å.

Outros métodos fornecem informações complementares:

- **Redes de recombinação de domínios.** Muitas proteínas eucarióticas contêm múltiplos domínios. Uma característica da evolução dos eucariotos é que um domínio pode aparecer em diferentes proteínas interagindo com outros domínios diferentes. Em alguns casos, as proteínas em um óperon de bactéria, que catalisam etapas sucessivas em uma via metabólica, estão fusionadas em uma única proteína multidomínio em eucariotos. Os domínios individuais da proteína eucariótica são homólogos às proteínas bacterianas. (Também são conhecidos exemplos de proteínas separadas em eucariotos e fusionadas em procariotos.)

  É possível criar uma rede ao se definir uma interação entre dois domínios protéicos sempre que homólogos dos dois domínios sejam observados em uma mesma proteína. Isso constitui evidência para algumas relações funcionais entre os domínios, mesmo em espécies onde esses domínios estejam presentes em proteínas distintas.

- **Padrões de co-expressão.** O agrupamento de dados de microarranjos identifica proteínas com padrões de expressão comuns. Elas podem apresentar a mesma distribuição tecidual ou terem a sua expressão aumentada ou reduzida em paralelo em diferentes estados fisiológicos. Esta é uma evidência indicativa de que essas proteínas compartilham relações funcionais. Na resposta à isoniazida pelo *M. tuberculosis* (página 328), genes do complexo de síntese de ácidos graxos (sistema FAS) têm sua expressão aumentada de maneira coordenada. Eles apresentam um arranjo similar a um óperon, e estas proteínas, de fato, formam um complexo físico. Por outro lado, a alquil-hidroperóxido redutase (AhpC) também tem aumento de expressão na presença de isoniazida. A AhpC age na atenuação dos efeitos do estresse oxidante. Não existe evidência de que ela interaja fisicamente com o sistema FAS, ou que ela medeie alguma transformação metabólica acoplada à síntese de ácidos graxos. Ela é um componente secundário da resposta à isoniazida.

- **Padrões de distribuição filogenética.** O **perfil filogenético** de uma proteína é o conjunto de organismos em que ela e seus homólogos são observados. Proteínas de um complexo estrutural comum ou em uma via comum são funcionalmente relacionadas e com co-evolução esperada. Dessa forma, proteínas que partilham um perfil filogenético provavelmente apresentam uma relação funcional, ou pelo menos uma origem subcelular comum. Não há necessidade de similaridade estrutural ou de seqüências entre as proteínas que partilham um padrão de distribuição filogenética. Uma característica importante deste método é que ele deriva informações sobre a função de uma proteína a partir de suas relações com outras proteínas *não-homólogas*.

Existem muitas maneiras de relacionar proteínas, incluindo interações físicas diretas proteína-proteína, complementaridade de duplo-híbrido, recombinação de domínios, padrões de co-expressão e perfis filogenéticos. Cada um destes métodos fornece um ponto de partida para a definição de uma rede de interações entre proteínas. As redes formadas pela combinação de cada conjunto de interações são diferentes, apesar de se sobreporem em maior ou menor proporção. Elas fornecem diferentes visões dos tipos de relações entre proteínas que podem existir em uma célula. É possível formar uma rede mais abrangente ao se combinar diferentes tipos de interações. Por exemplo, o banco de dados DIP

(http://dip.doe-mbi.ucla.edu/) é uma coletânea de interações proteína-proteína determinadas experimentalmente. Ele contém dados sobre 56.048 interações entre 19.344 proteínas de 154 organismos diferentes.

A Gravura X mostra uma porção da rede de interações de proteínas de levedura, baseada no conjunto de proteínas que foram observadas formando complexos em estruturas resolvidas.

Recursos na *web*: Bancos de dados de interações

Intact: Um banco de dados de interações moleculares
http://www.ebi.ac.uk/intact/

DIP: Database of Interacting Proteins
http://dip.doe-mbi.ucla.edu/

MIPS: Banco de dados abrangente do genoma de levedura
http://mips.gsf.de/

BIND: Biomolecular Interaction Network Database
http://www.bind.ca/

MINT: A Molecular Interactions Database
http://cbm.bio.uniroma2.it/mint/

SPiD: *Subtilis* Protein interaction Database
http://genome.jouy.inra.fr/cgi-bin/spid/index.cgi

GRID: General Repository for Interaction Datasets
http://biodata.mshri.on.ca/grid/servlet/Index

PathCalling: Interações proteína-proteína em *S. cerevisiae*
http://portal.curagen.com/pathcalling_portal/index.htm

HPID: The Human Protein Interaction Database
http://www.hpid.org/

Estudo de Caso 6.3: Componentes da formação do primossomo em *Bacillus subtilis*

A primeira etapa na replicação do DNA em *Bacillus subtilis* é a ligação de proteínas de iniciação a seqüências específicas de DNA que atuam como origens de replicação. Após essa ligação, ocorre o recrutamento de um complexo de nucleoproteínas chamado de primossomo. Um dos principais componentes do primossomo é DnaC, uma helicase replicativa hexamérica.*

Acredita-se que as etapas deste processo incluam:

1. Ligação de uma proteína iniciadora, DnaA ou PriA, a uma seqüência de DNA fita simples apropriada.

* É importante ressaltar que a nomenclatura destas proteínas é diferente em *E. coli* e *B. subtilis*.

Estudo de Caso 6.3 (*continuação*)

2. Outras proteínas – DnaB, DnaC e DnaI – são recrutadas. DnaB e DnaI regulam a atividade da DnaC.
3. DnaC é disposta em uma cadeia de DNA fita simples, formando a estrutura hexamérica.
4. DnaG é recrutada para dar início à síntese de DNA.

Cientistas do Institut National de la Récherche Agronomique mantêm um banco de dados da rede de interações de proteínas de *B. subtilis** (`http://genome.jouy.inra.fr/cgi-bin/spid/index.cgi`).

A Figura 6.6 mostra um pequeno fragmento da rede, limitada às proteínas vizinhas da DnaC.

O sítio na *web* é interativo: clicando em um nó, ocorre a *adição* das proteínas que interagem com o nó no grafo ou a *substituição* do grafo por outro centralizado na proteína selecionada. Com a adição de mais proteínas interagentes, é possível observar uma vizinhança mais extensa da DnaC. Com a substituição do grafo, é possível se deslocar através da rede (ver *Weblema* 6.5).

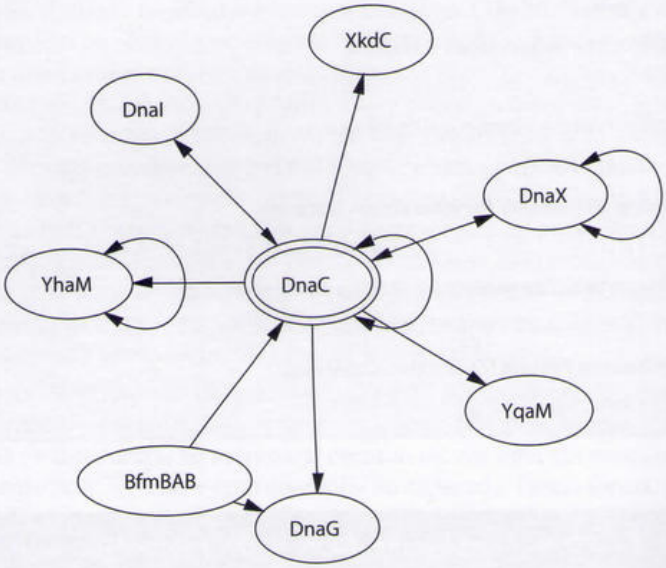


Figura 6.6 DnaC e as proteínas com as quais interage diretamente. As setas conectam os pares de proteína no sentido de "isca" para "predador"; setas bidirecionais indicam casos onde a interação foi detectada em experimentos recíprocos. No sítio da *web* original, as setas são coloridas de acordo com a natureza da evidência da interação. Reproduzida com permissão.

* Hoebeke, M., Chiapello, H., Noirot, P. & Bessiéres, P. (2001), SPiD: a subtilis protein interaction database. *Bioinformatics*, **17**,1209-1212; Noirot-Gros, M. F., Dervyn, E., Wu, L. J., Mervelet, P., Errington, J., Erlich, S. D. & Noirot, P. (2002), An expanded view of the bacterial DNA replication. *Proc. Natl. Acad. Sci. USA*, **99**, 8342-8347.

Redes reguladoras

As interações de controle individuais são organizadas em cascatas de transdução de sinal lineares e organizadas em redes de controle. Redes reguladoras são observadas em todos os processos vivos.

Qualquer ação reguladora demanda (1) um estímulo, (2) a transmissão de um sinal a um alvo, (3) uma resposta e (4) um mecanismo de "reinicialização" para restaurar o estado de repouso (ver Figura 6.7). Muitas ações reguladoras são mediadas por complexos proteína-proteína. Complexos transitórios são comuns na regulação, pois a dissociação fornece um mecanismo natural de reinicialização.

Alguns estímulos têm a sua origem em programas genéticos. Alguns eventos reguladores são respostas à concentração interna de metabólitos em um dado momento. Outros se originam no meio extracelular; o sinal é detectado por receptores de superfície e transmitido através da membrana para um alvo intracelular.

Dois componentes de redes reguladoras são (1) a **rede de transdução de sinais** e (2) a **rede de controle transcricional**. A rede de transdução de sinais exerce controle *"no campo de batalha"* por uma variedade de mecanismos, tais como inibidores; dimerização, mudanças conformacionais induzidas por ligante, incluindo mas não se limitando aos efeitos alostéricos; trocas GDP-GTP ou "interruptores" de cinase-fosforilase; e taxas diferenciais de reciclagem e conversão de moléculas. Este componente é de ação rápida, na escala de tempo de subsegundos. A rede de controle transcricional exerce controle *"no quartel-general"*, por meio do controle da expressão gênica. Este componente é mais lento, agindo em uma escala de tempo de minutos.

Características gerais de todas as vias de controle:

- Um único sinal pode disparar uma única resposta ou muitas respostas.

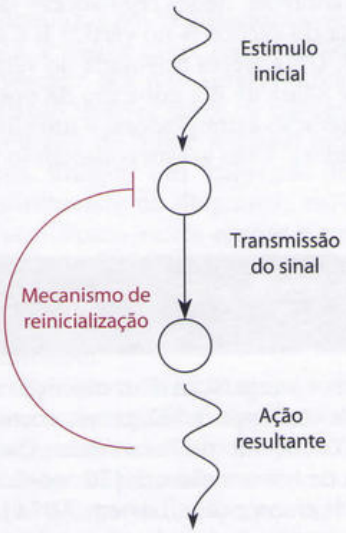



Figura 6.7 Etapas elementares em uma rede reguladora. Um estímulo inicial é recebido por um nó, o qual transmite um sinal para os nós subseqüentes, causando uma ação resultante. Esta resposta é seguida pela reinicialização dos nós situados acima para que retornem ao seu estado inativo. A combinação destes diagramas elementares produz as complexas redes reguladoras na biologia.

- Uma única resposta pode ser controlada por um único sinal ou ser influenciada por muitos sinais.
- Cada resposta pode ser estimuladora – aumento da atividade – ou inibidora – diminuição da atividade.
- A transmissão de sinais pode diminuir um estímulo ou amplificá-lo.

Há muitas formas de se atingir complexidade, e as células aproveitam todas as oportunidades de empregá-las.

Estruturas de redes reguladoras

Pense nas redes de controle, ou reguladoras, como agrupamentos de *atividades*. Embora mediadas em parte por uniões físicas de macromoléculas – complexos proteína-proteína e proteína-ácido nucléico – as redes reguladoras:

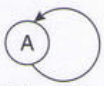
(1) **Tendem a ser unidirecionais.** Um ativador de transcrição pode estimular a expressão de uma enzima metabólica, mas a enzima pode não estar envolvida diretamente na regulação da expressão do fator de transcrição (ver página 102 para uma discussão sobre o controle da via de síntese do triptofano em *E. coli*).

interação estimuladora

(2) **Possuem um componente lógico.** Descrever a conectividade de uma rede reguladora não é o suficiente. Qualquer ação reguladora pode estimular ou reprimir a atividade do seu alvo. Se duas interações se combinam para ativar um alvo, a ativação pode exigir os *dois* estímulos (E lógico), ou *qualquer* um dos estímulos pode ser o suficiente (OU lógico).

interação inibidora

(3) **Produzem padrões dinâmicos.** Os sinais podem produzir combinações de efeitos em cursos de tempo específicos. A regulação do ciclo celular é um exemplo clássico.



interação auto-reguladora

Uma rede reguladora pode ser descrita por um grafo, no qual as arestas indicam as etapas na via de controle. Redes reguladoras são grafos orientados (ver página 335): a influência do vértice A no vértice B é expressa por uma aresta orientada ligando A e B. Uma aresta orientada do vértice A para o vértice B é chamada de **conexão de saída** de A e **conexão de entrada** em B. Geralmente, uma seta indica uma interação estimuladora, e um símbolo em forma de T indica uma interação inibidora. Uma aresta conectando um vértice a ele mesmo

interação recíproca

Estudo de Caso 6.4: Arquitetura e dinâmica da rede reguladora genética de *Saccharomyces cerevisiae

Um estudo recente sobre a regulação de transcrição em levedura levou em consideração uma rede contendo 3.562 genes, correspondendo a cerca de metade do proteoma conhecido de *S. cerevisiae*. Dentre esses genes, 142 codificam reguladores de transcrição e 3.420 codificam genes-alvo específicos dos reguladores de transcrição. Existem 7.074 interações reguladoras conhecidas entre esses genes, incluindo efeitos dos reguladores uns sobre os outros e de genes reguladores sobre genes-alvo não-reguladores.

* Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M. Teichmann, S. A. & Gerstein, M. B. (2004), Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**,308-312.

→

A análise da arquitetura geral da rede revela que:

- A distribuição de conexões de entrada para genes-alvo tem o valor médio de 2,1 e é distribuída exponencialmente. A maioria dos genes-alvo recebe sinais de entrada diretamente de cerca de dois reguladores de transcrição. A probabilidade de um gene ser controlado por k reguladores de transcrição, onde $k = 1, 2, ...$, é proporcional a $e^{-}\alpha^{k}$, com $\alpha = 0{,}8$.
- A distribuição das conexões de saída tem valor médio de 49,8 e obedece a uma lei de potência. A probabilidade de um dado regulador de transcrição controlar k genes é proporcional a $k^{-}\beta$, com $\beta = 0{,}6$. Comportamentos regidos pela lei de potência são comuns em redes e caracterizam topologias em que poucos nós – os "roteadores" – possuem muitas conexões e muitos nós possuem poucas (ver página 337). Em redes reguladoras, os roteadores tendem a estar localizados bem no topo da rede, formando importantes centros de controle de longo alcance.
- O número médio de nós intermediários em um caminho mínimo entre um regulador de transcrição e um gene-alvo é 4,7. O número máximo de nós intermediários em um caminho entre dois nós é 12.
- O coeficiente de agrupamento de um nó é uma medida do grau de conectividade local dentro de uma rede. Se todos os vizinhos de um nó estiverem conectados uns aos outros, o coeficiente de agrupamento deste nó é 1. Se nenhum par de vizinhos de um nó estiver conectado um ao outro, o coeficiente de agrupamento do nó é 0. O coeficiente de agrupamento médio, considerando a média sobre todos os nós, é uma medida da densidade da rede inteira. Para a rede de regulação transcricional de levedura, o coeficiente de agrupamento médio é 0,11.

A Figura 6.8 é uma representação esquemática de um fragmento desta rede, indicando, de forma simplificada, algumas das características gerais. Os nós estão divididos em **reguladores de transcrição**, representados como círculos, e **genes-alvo**, representados como quadrados. Os genes-alvo se distinguem por não apresentarem conexões de saída. Existe uma inter-regulação intensa entre os fatores de transcrição, com uma densidade muito maior do que é possível mostrar de maneira inteligível neste diagrama. Imagine um "caldeirão de sopa" de fatores de transcrição, na área sombreada do diagrama, enviando sinais para os genes-alvo. Essa área sombreada indica apenas o agrupamento *lógico* dos reguladores de transcrição. Não existem indicações sobre suas localizações físicas: na verdade, os reguladores de transcrição interagem com a cadeia de DNA e quase nunca interagem fisicamente com as proteínas cuja expressão eles controlam.

Cada regulador da transcrição influencia diretamente cerca de 50 genes, em média, embora, como outras redes do tipo "mundo pequeno", cuja distribuição de conectividades segue a lei de potência, a distribuição é bastante distorcida – alguns "roteadores" possuem um elevado número de conexões de saída, e a maioria dos nós possui muito poucas. Poucas das conexões inter-reguladoras entre fatores de transcrição são mostradas em vermelho. Em cerca de 10% dos casos, dois vizinhos do mesmo fator de transcrição interagem entre si. Uma via a partir de um gene regulador (círculo preto) até um receptor final (quadrado preto), através de cinco nós intermediários, é mostrada em preto. Os nós intermediários são outros reguladores

Estudo de Caso 6.4 (*continuação*)

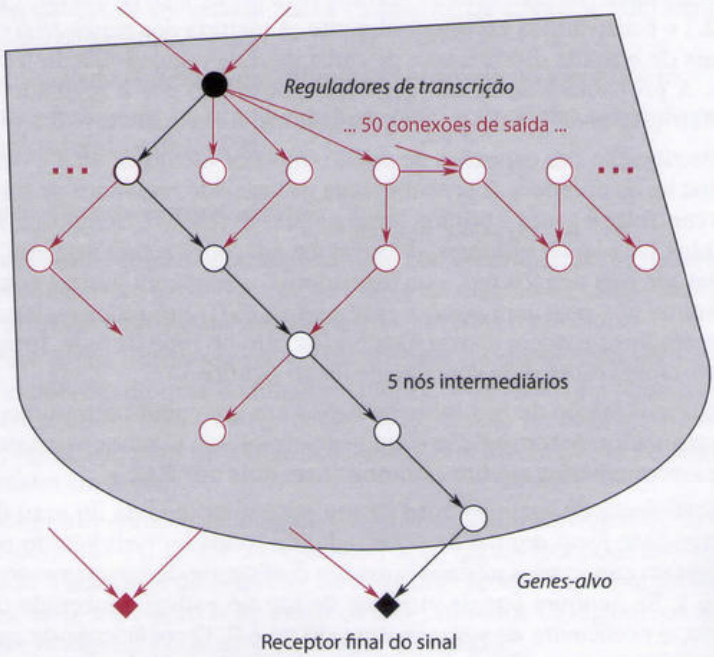


Figura 6.8 Representação simplificada ilustrando algumas das características de um segmento "médio" das vias na rede de interações em levedura. Reguladores de transcrição estão representados como círculos. Genes-alvo estão representados como quadrados. Um regulador de transcrição geralmente possui influência direta sobre cerca de 50 genes, indicada pelas múltiplas conexões a partir do círculo preto em direção aos círculos da linha abaixo dele. Aproximadamente um em cada dez vizinhos de qualquer nó está conectado a outro vizinho, indicado pela seta horizontal na segunda linha. O receptor final do sinal se encontra no final de uma via, que tipicamente contém cerca de cinco nós intermediários (representados em preto). Este gene-alvo final recebe, em média, duas conexões de entrada. Este diagrama mostra apenas uma pequena porção de uma rede que é, na verdade, bastante densa.

de transcrição, com conexões dentro da via mostrada em preto e também fora dessa via. Mesmo o fator de transcrição, mostrado como a origem da via, recebe conexões de entrada. Embora seja possível identificar os genes-alvo pela ausência de conexões de saída, é mais complicado identificar os iniciadores fundamentais de cascatas de sinalização.

O receptor final é um gene-alvo que recebe sinais reguladores, mas ele próprio não apresenta conexões de saída. Espera-se que esse gene-alvo receba, em média, um segundo sinal de controle. O nó-alvo em preto recebe conexões de entrada através das setas em preto, ao longo do caminho selecionado, e através das setas em vermelho, indicando uma segunda conexão de entrada. Com certeza, essa segunda conexão pode se originar em um caminho que compartilha alguns nós com o caminho representado em preto, incluindo outras rotas a partir do círculo preto inicial.

A densa floresta de vias adicionais, da qual este fragmento foi extraído, não é mostrada. Alguns cálculos básicos: há cerca de 3.500 nós, cada um recebendo, em média, duas conexões de entrada. Há cerca de 140 fatores de transcrição, fazendo uma média de 50 conexões de saída. O número de conexões de entrada deve ser igual ao número de conexões de saída e, de fato, $3.500 \times 2 = 140 \times 50 = 7.000$.

Dada esta complexidade, é difícil ilustrar segmentos maiores da rede em mais detalhes do que a forma simplificada mostrada na Figura 6.8. Entretanto, as dissecações da rede reguladora de levedura e de outros organismos definiram alguns motivos recorrentes que servem como pilares para a construção de redes. Estes elementos podem ser considerados as "estruturas secundárias" das arquiteturas de redes (ver Quadro Motivos comuns em redes biológicas de controle, página 355).

A alta proporção de interações com relação a reguladores de transcrição significa que não podemos esperar a associação de moléculas reguladoras individuais com atividades únicas e dedicadas (como podemos fazer para a maioria das enzimas metabólicas). Ao contrário, a atividade das redes envolve as ações coordenadas de muitas moléculas reguladoras individuais.

A rede alcança versatilidade e reatividade pela reconfiguração das suas atividades. Isso pode ser observado comparando as mudanças nas atividades das redes que controlam padrões de expressão gênica em levedura em diferentes condições fisiológicas do organismo: ciclo celular, esporulação, desvio diáuxico (mudança do metabolismo de fermentação anaeróbia para a respiração aeróbia conforme aumentam os níveis de O_2), dano ao DNA e resposta ao estresse. O ciclo celular e a esporulação envolvem o desencadeamento de programas endógenos de expressão gênica; os demais exemplos são respostas a mudanças ambientais.

Diferentes estados são caracterizados pelas similaridades e diferenças no padrão de expressão gênica e pelos componentes da rede reguladora que estão ativos. Existe um desvio considerável na expressão dos *genes-alvo*. Cerca de um quarto dos genes-alvo são especializados para algum estado fisiológico específico. Ou seja, do total de 3.420 genes-alvo, a expressão de cerca da metade (1.514) não se altera em estados fisiológicos distintos. Dos 1.906 genes que apresentam níveis de expressão alterados em diferentes estados, cerca da metade (803) são especializados para um único estado fisiológico.

Em contraste, diferentes estados apresentam maior sobreposição no uso dos *reguladores de transcrição*. Por exemplo, no controle do ciclo celular, 280 genes-alvo (8%) têm expressão diferencialmente regulada por 70 (49%) dos reguladores de transcrição. Existe claramente um menor grau de especialização do que o observado nos genes-alvo. Em geral, metade dos fatores de transcrição está ativa em pelo menos três dos cinco regimes fisiológicos. No entanto, contrastando com a grande sobreposição no uso dos reguladores de transcrição (os nós), a sobreposição de atividades dentro da rede (as conexões) é relativamente baixa. Diferentes componentes da rede de interações organizam os diferentes padrões de expressão gênica em estados diferentes.

Estudo de Caso 6.4 (*continuação*)

Enquanto diferentes estados fisiológicos são caracterizados por substituições de conjuntos diferentes de proteínas sintetizadas, a rede reguladora utiliza muito da mesma estrutura, mas reconfigura seu padrão de atividade. Pense nos fatores de transcrição como *hardware* e nas conexões como *software* reprogramáveis. As moléculas não se alteram, mas as suas interações, sim: em estados diferentes, muitos reguladores de transcrição mudam a maior parte, ou uma parte substancial, das suas interações. Em particular, o conjunto de reguladores de transcrição que formam os roteadores da rede – aqueles com muitas conexões de saída que formam centros de controle – não é uma característica constante do sistema. Alguns roteadores são comuns a todos os estados, mas outros se apresentam para tomar o controle em regimes fisiológicos diferentes. O resultado da reconfiguração de atividade é que mais da metade das interações reguladoras são *únicas* a cada estado fisiológico.

O efeito das mudanças nos padrões de interações ativas é a alteração das características topológicas da rede em diferentes estados. Por exemplo, sob condições de pânico – dano ao DNA e estresse – o número médio de genes sob o controle de reguladores de transcrição individuais aumenta, o comprimento médio do caminho mínimo entre o regulador e seu alvo diminui, e o agrupamento se torna menos denso (ou seja, existe menos inter-regulação entre os fatores de transcrição). Isso pode ser entendido em termos da necessidade de uma mobilização rápida e geral – o equivalente de gritar "Anda! Anda! Anda!" através de um rádio em uma situação de emergência. Circunstâncias normais – como o controle do ciclo celular, por exemplo – possibilitam um estado regulador mais preciso, permitindo o controle mais fino sobre o curso temporal dos padrões de expressão. No controle do ciclo celular e da esporulação, existe uma inter-regulação muito mais densa entre os fatores de transcrição e comprimentos de caminhos mínimos mais longos entre os reguladores e os genes-alvo.

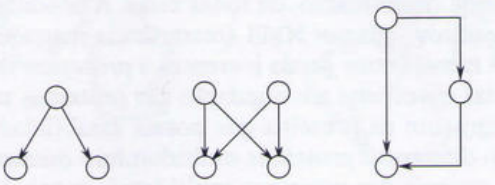
Diferentes estados fisiológicos também diferem no uso dos motivos comuns – forquilha, disperso e "*one-two punch*" (ver Quadro Motivos comuns em redes biológicas de controle). Forquilhas são mais utilizadas em condições de estresse, desvios diáuxicos e dano ao DNA. Elas são apropriadas à necessidade de respostas rápidas. As exigências para a formação de intermediários causariam um atraso na resposta. Ao contrário, o motivo "*one-two punch*" é mais comum no controle do ciclo celular. Isso é consistente com a necessidade de um sinal de um estágio ser estabilizado antes de a célula entrar no próximo estágio.

Grande parte da evolução se direciona para uma maior especialização. O olho humano é um exemplo clássico. O olho é uma estrutura complexa e muito especializada (características que já foram utilizadas como evidências *contra* a teoria de Darwin). Muitas vias evolucionárias mostram o balanço entre a adaptação especializada e a adaptabilidade generalizada.

As redes reguladoras são uma exceção. A evolução produziu estruturas que são *tanto* especializadas *quanto* versáteis. A capacidade de reconfiguração de redes reguladoras permite que elas respondam de forma robusta a alterações nas condições, criando estruturas diferentes, todas especializadas nas condições que as tornam ativas.

Motivos comuns em redes biológicas de controle

Dentro da alta complexidade de redes reguladoras típicas, certos padrões comuns aparecem com freqüência. Na arquitetura de redes, estes motivos formam os pilares de um nível superior de organização. Shen-Orr, Milo, Mangan e Alon* descreveram exemplos que incluem os motivos *forquilha, disperso* e *"one-two punch"* (um termo emprestado do boxe):



A **forquilha**, também chamada de motivo de uma única entrada, transmite um único sinal de entrada para duas saídas. Forquilhas sucessivas, ou forquilhas com graus maiores de ramificação, são um meio eficaz de se ativar grandes conjuntos de genes a partir de um único impulso. Generalizações da forquilha binária podem incluir mais genes sob um controle comum (mais ramos na forquilha) e auto-regulação do nó de controle. As forquilhas podem alcançar a mobilização geral. Além disso, se os genes regulados apresentarem diferentes limiares de ativação, a dinâmica de geração do sinal pode causar um padrão temporal de iniciação sucessiva da expressão de diferentes genes.

A configuração **dispersa**, também chamada de motivo de entradas múltiplas, pode funcionar como uma operação lógica OU: os dois genes-alvo sob seu controle podem ser ativados se *uma* das duas conexões de entrada estiver ativa. Generalizações do padrão de dispersão quadrado mostrado podem conter números diferentes de nós em qualquer uma das camadas. Note que os padrões dispersos são sobreposições de forquilhas.

O motivo ***"one-two punch"***, também chamado de "laço de alimentação positiva", afeta a saída diretamente pela conexão vertical e, indireta e subseqüentemente, pelas conexões intermediárias. Este motivo pode apresentar um comportamento temporal interessante se a ativação do alvo exigir a entrada simultânea dos caminhos diretos e indiretos (E lógico). Como a formação do intermediário demanda tempo, o sinal direto chegará ao alvo antes do sinal indireto. Dessa forma, um sinal de entrada de curta ação sobre o complexo não ativará o alvo – pois no momento em que o sinal intermediário se formar, o sinal direto já não estará ativo. O sistema pode então filtrar os estímulos transitórios nos sinais de entrada. De forma complementar, o estado ativo do sistema ainda pode ser desligado rapidamente em resposta à diminuição ou retirada do estímulo externo.

Um exemplo deste tipo de mecanismo de controle são as ações de um motorista em resposta a um semáforo. A reação a uma luz amarela seguida da luz verde deveria ser uma aceleração cuidadosa. A resposta a uma luz vermelha deve ser uma parada imediata. O controle sobre o hábito de beber infelizmente *não permite* mostrar este tipo de comportamento: a primeira dose leva ao consumo de mais doses, não havendo nenhuma maneira rápida para ficar sóbrio!

* Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002), Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64-68.

indica auto-regulação. Uma seta bidirecional indica estimulação recíproca de dois nós; note que isso *não* é o mesmo que uma aresta não-orientada.

Biologia estrutural de redes reguladoras

Toda interação reguladora envolve uma ou mais proteínas e ácidos nucléicos. Exemplos de mecanismos reguladores incluem a ligação entre uma proteína e um ligante, a ocorrência de modificações químicas como fosforilação e desfosforilação, mudanças na conformação, ou todas estas. A cristalografia por difração de raios X e a espectroscopia por NMR (ressonância magnética nuclear) elucidaram alguns dos mecanismos gerais inerentes a processos de controle.

Muitas moléculas envolvidas em regulação são proteínas multidomínio. Um domínio é um segmento de proteína que possui estabilidade independente e pode aparecer em diferentes proteínas multidomínio mediante recombinação evolucionária. A maioria das proteínas multidomínio contém uma seqüência linear de domínios, cada um dos quais é relativamente livre para interagir com outras moléculas. A formação de uma proteína a partir de domínios permite, portanto, a união, em uma única molécula, de um conjunto de funções. A "mistura e combinação" de domínios oferece à evolução acesso a uma ampla variedade de combinações funcionais (ver Figura 2.3, página 131).

Uma característica importante das proteínas reguladoras é o reconhecimento. Um **domínio de interação** é uma parte de uma proteína que confere especificidade na ligação com outro domínio (ou subunidade). As proteínas reguladoras contêm um número limitado de tipos de domínios de interação, os quais divergiram para formar grandes famílias com especificidades individuais diferentes. Por exemplo, o genoma humano contém 115 domínios SH2 e 253 domínios SH3. (Domínios Src-homólogos SH2 e SH3 são assim nomeados por causa de suas homologias com os domínios da família src de tirosina cinases citoplasmáticas.) Muitos domínios de interação individuais interagem até mesmo com diferentes proteínas e/ou ligantes quando atuam em etapas sucessivas de cascatas de controle. As interações iniciais podem também levar ao recrutamento de proteínas adicionais para formar grandes complexos reguladores.

Muitos domínios de interação são sensíveis ao estado de modificação pós-traducional de seus ligantes, por exemplo, ligando-se preferencialmente a estados de um ligante no qual resíduos específicos de tirosina, serina ou treonina estão fosforilados. Estas e outras modificações pós-traducionais funcionam como "interruptores" reversíveis, iniciando ou interrompendo/reiniciando uma cascata de sinalização.

A formação de complexos proteína-proteína permite a uma célula detectar uma molécula sinalizadora no meio externo e comunicar a sua detecção ao interior celular, sem que a própria molécula sinalizadora precise ser internalizada pela célula. Muitos receptores usam um mecanismo engenhoso de dimerização: o receptor possui domínios externos, transmembrana e internos. Um ligante externo se liga a *duas* moléculas de receptor. A justaposição dos domínios externos também aproxima os domínios internos, pois esses dois domínios estão "amarrados" pelo domínio transmembrana. A interação entre os domínios internos provoca uma mudança conformacional que leva à ativação de um processo, tal como a fosforilação de uma proteína. Isso permite a inicialização de uma cascata de transdução de sinais que pode amplificar o estímulo original.

A Figura 6.9 mostra tipos de complexos de domínios de interação com ligantes, incluindo a ligação de peptídeos (que podem estar ligados a proteínas), complexos proteína-proteína, formação de dímeros extracelulares após a ligação de um hormônio e complexos proteína-ácido nucléico.

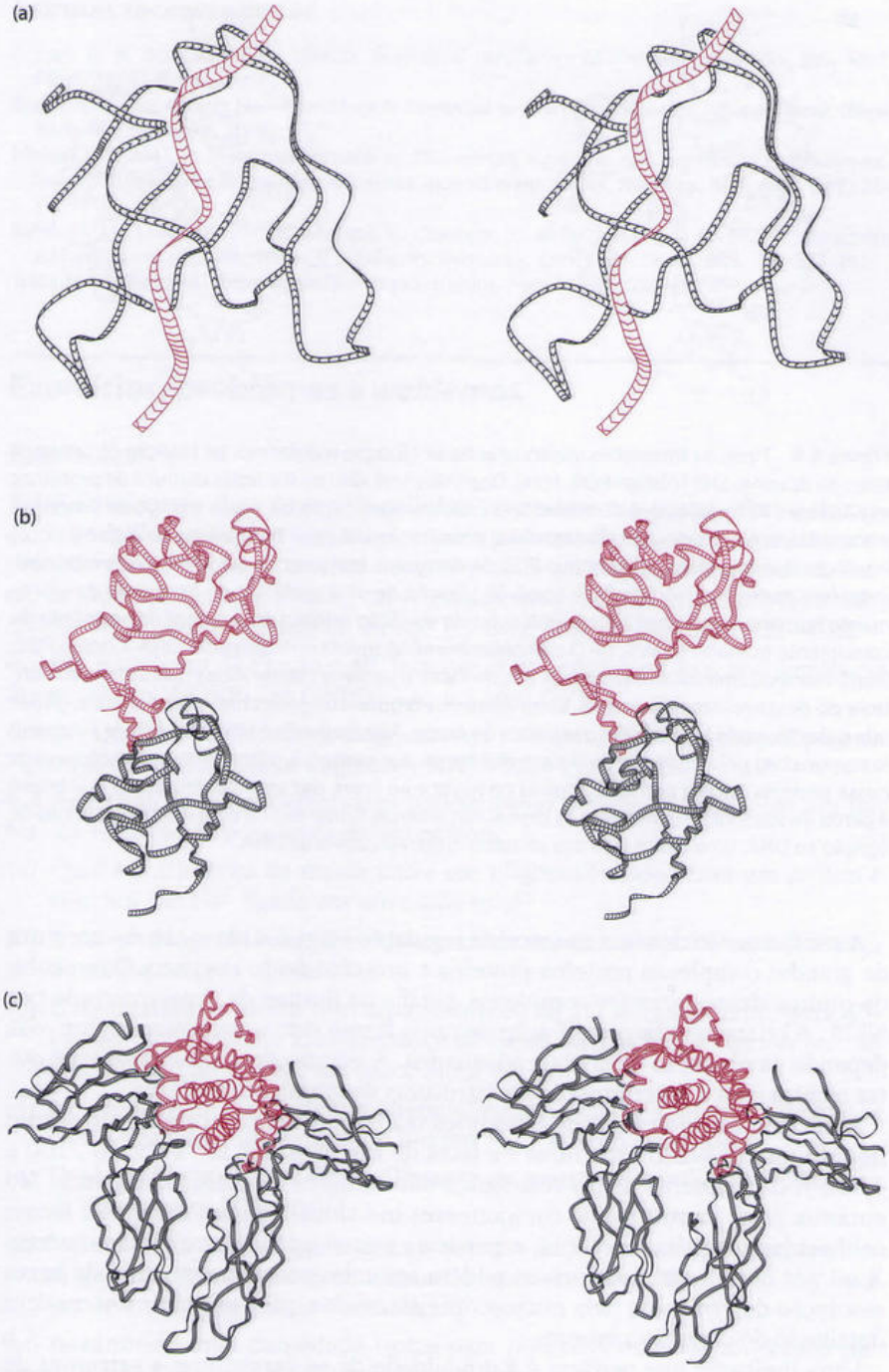


Figura 6.9 *Continua.*

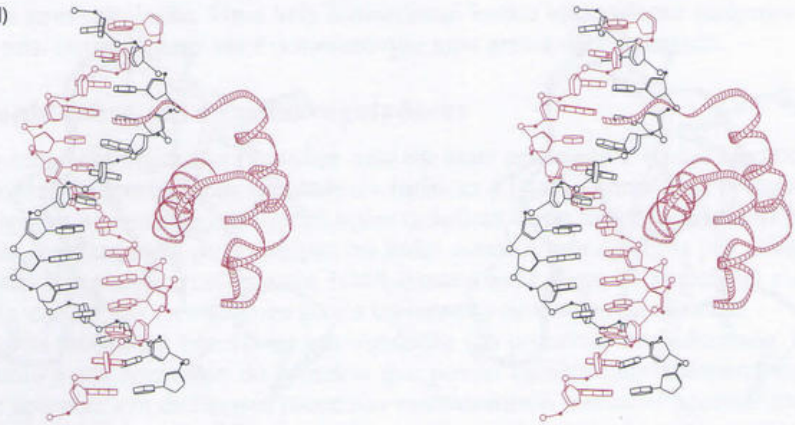


Figura 6.9 Tipos de interações observadas na sinalização reguladora. (a) Ligação de um peptídeo ao domínio SH3 [código PDB: 1CKA]. Domínios SH3 são constituintes comuns de proteínas reguladoras. As funções dos domínios SH3 incluem transdução de sinais, tráfego de proteínas e vesículas, organização do citoesqueleto, polarização celular e biossíntese de organelas. (b) Interação domínio-domínio: domínio PDZ da sintropina (em preto) e óxido nítrico-sintase neuronal (em vermelho) [código PDB: 1QAV]. (c) Ligação de uma molécula de hormônio de crescimento humano (vermelho) a duas moléculas do domínio externo do receptor de hormônio de crescimento humano (preto). (d) O complexo homeodomínio *antennapedia*-DNA [Código PDB: 9ANT]. Homeodomínios são proteínas eucarióticas altamente conservadas que atuam no controle do desenvolvimento animal. Estes domínios regulam os genes homeóticos, isto é, genes que especificam as localizações das partes do corpo. *Antennapedia* é uma proteína de *Drosophila* responsável pela iniciação do desenvolvimento das pernas. A primeira mutação observada nessa proteína origina pernas ectópicas no lugar, e ao invés, das antenas. Mutações que levam à perda de sua função convertem as pernas em antenas. Como ocorre com muitas proteínas de ligação ao DNA, uma hélice α se liga ao sulco maior da cadeia de DNA.

A compreensão dos mecanismos de regulação exigirá a obtenção da estrutura de grandes complexos proteína-proteína e proteína-ácido nucléico. O tamanho de muitos desses grandes complexos desafia os limites da espectroscopia por NMR. A cristalografia por difração de raios X tem tido um sucesso maior, mas depende da obtenção de cristais adequados. A microscopia crioeletrônica é outra técnica para a determinação de estruturas de grandes complexos.

A microscopia eletrônica de espécimes realizada à temperatura do nitrogênio líquido tem revelado estruturas na faixa de r.m.m = 5×10^5 a 4×10^8, 100 a 1.500 Å de diâmetro. Estes resultados não atingem a resolução atômica. No entanto, se as estruturas de componentes individuais de um complexo forem conhecidas com alta resolução, a partir de cristalografia por difração de raios X ou por NMR, essas estruturas podem ser sobrepostas à estrutura de baixa resolução determinada pela microscopia eletrônica, para produzir um modelo detalhado do complexo inteiro.

Uma limitação que perdura é a dificuldade de se determinar a estrutura de complexos transitórios ou de sistemas que apresentam mudanças conformacionais significativas durante a formação do complexo. Esta situação se repete em muitas áreas da biologia molecular atual: estamos compreendendo melhor estruturas estáticas, mas esperando o desenvolvimento de métodos para tratar os aspectos dinâmicos.

Leituras recomendadas

Albert, R. & Barabàsi, A.-L. (2002), Statistical mechanics of complex networks, *Rev. Mod. Phys.*, **74**, 47-97.

Barabàsi, A.-L., *Linked: How Everything Is Connected to Everything Else and What It means* (New York: Plume Books, 2003)

Ideker, T. (2004), A systems approach to discovering signaling and regulatory pathways–or, how to digest large interaction networks into relevant pieces, *Adv. Exp. Med. Biol.*, **547**, 21-30.

Babu, M. M., Luscombe, N. M, Aravind, L., Gcrstein, M. & Teichmann, S. A., (2004), Structure and evolution of transcriptional regulatory networks, *Curr. Opin. Struct. Biol.*, **14**, 283-291.

Tyers M. & Mann M., From genomics to proteomics, *Nature*, **422**, 193-197.

Exercícios, problemas e *weblemas*

Exercícios

6.1 A lisozima da clara de ovo de galinhas tem massa molecular relativa de cerca de 14.300. Se a espectrometria de massa pode medir variações na massa com precisão de 0,01%, as alterações a seguir podem ser distinguidas, com confiança, da proteína não-modificada? (a) Acetilação da porção N-terminal? (b) Fosforilação de um único resíduo de serina? (c) Uma única substituição Lis→Gln?

6.2 Em fotocópias da Figura 6.4b, indique as posições dos picos se a seqüência fosse: (a) MNLVQVR, (b) GNLQVVR, (c) MNLQVVG.

6.3 (a) Qual é a seqüência do fragmento y_6 na Figura 6.4b? (b) A qual pico da Figura 6.4b corresponde o fragmento NH_3^+ – LQVVR-COOH?

6.4 Amostras de oligonucleotídeos podem variar através da ligação de um íon Na^+ ou K^+ ao fosfato ao invés de um próton.

(a) Qual é a diferença de massa entre um oligonucleotídeo com um próton e com um íon Na^+ ligado em um único local?

(b) Qual mudança de base tem a diferença de massa mais próxima à diferença de massa entre H^+ e Na^+?

(c) Uma medida de massa com uma resolução de 1D seria suficiente para distinguir corretamente esta mudança de bases da ligação de um íon Na^+ ao invés de um próton em um único sítio?

(d) No espectro de massa de um oligonucleotídeo, qual é a diferença de massa com a ligação de um próton ou um íon Mg^{2+} em um único sítio?

(e) Qual mudança de base tem a diferença de massa mais próxima à diferença de massa entre H^+ e Mg^{2+}?

(f) Uma medida de massa com uma resolução de 1D seria suficiente para distinguir corretamente esta mudança de bases da ligação de um íon Mg^{2+} em vez de um próton em um único sítio?

6.5 Assumindo uma densidade típica para polimorfismos de um único nucleotídeo (SNPs) como 1 SNP/5 kpb no genoma humano, e apenas duas bases possíveis observadas na posição de qualquer SNP, quantas seqüências se pode esperar encontrar em uma população, em uma região de 100 kpb, se a recombinação for comum em todas as posições dessa região? Se apenas três das possíveis combinações de SNPs – ou seja, três haplótipos – forem observadas, que fração das seqüências possíveis esses haplótipos representam?

6.6 Para quais métodos de determinação de proteínas interagentes (páginas 345-346) (a) uma das proteínas precisa ser purificada antes de a interação ser testada? (b) as duas proteínas precisam ser purificadas antes de a interação ser testada?

6.7 No primeiro grafo na margem da página 335:

(a) Indique dois vértices, tais que se uma aresta for adicionada entre eles, pelo menos um vértice terá exatamente dois vizinhos. (Note que duas arestas podem se cruzar sem adicionar um novo vértice no ponto de sua intersecção.)

(b) Indique dois vértices tais que, se for adicionada uma aresta entre eles no grafo original, o grafo se mantenha uma árvore.

(c) Indique dois vértices (nenhum deles podendo ser V_1) tais que, se for adicionada uma aresta entre eles no grafo original, o grafo deixe de ser uma árvore.

(d) Indique dois vértices tais que, se for adicionada uma aresta entre eles no grafo original, existam caminhos alternativos de comprimentos 3 e 4 entre V_1 e V_5 sem vértices repetidos. (Na determinação do comprimento de um caminho, você tem de incluir os vértices inicial e final. Um caminho de comprimento 3 entre V_1 e V_5 contém um vértice intermediário.)

(e) Indique dois vértices tais que, se for adicionada uma aresta entre eles no grafo original, exista exatamente um caminho entre V_1 e V_3, sem vértices repetidos e com comprimento igual a 4.

6.8 Nos exemplos de grafos no final da página 335, (a) Quais são grafos orientados? (b) Quais são grafos rotulados? (c) Em cada exemplo, qual é o conjunto de nós? (d) Em cada exemplo, qual é o conjunto de arestas?

6.9 Em uma interface proteína-proteína típica, com área igual a 1.700 $Å^2$: (a) Quantas ligações de hidrogênio intermolecular espera-se que sejam formadas? (b) Quantas moléculas fixas de água espera-se encontrar nessa interface? (c) Se toda a superfície da interface fosse hidrofóbica, qual seria a sua estimativa da contribuição dessa superfície para a energia livre de estabilização?

6.10 Do fragmento da rede de interação de proteínas de *B. subtilis* mostrado na Figura 6.6, qual é o coeficiente de agrupamento da proteína DnaC? O coeficiente de agrupamento de um nó em um grafo é definido da seguinte forma: suponha que um nó contenha k vizinhos. O total de conexões possíveis entre vizinhos é $k(k-1)/2$. O coeficiente de agrupamento é o número de conexões observadas entre os vizinhos, dividido pelo número máximo de potenciais conexões entre os vizinhos.

6.11 No metrô de Londres: (a) Qual é o menor caminho entre as estações Moorgate e Embankment? Note que, considerado como um grafo, o menor caminho entre dois nós é o caminho com menor número de nós intermediários, e não o caminho com menor tempo de viagem ou com menos troca de trens. (b) Qual é o ciclo mais curto que contém as estações King's Cross, Holborn e Oxford Circus? (c) Se os vizinhos de uma estação são outras estações que podem ser alcançadas sem a passagem por estações intermediárias, qual é o coeficiente de agrupamento da estação Oxford Circus? (Ver Exercício 6.10 para a definição do coeficiente de agrupamento.)

6.12 No metrô de Londres: (a) Qual é o maior comprimento de caminho entre duas estações quaisquer? Ou seja, para quais duas estações a menor viagem

entre elas envolve o número máximo de paradas intermediárias? (b) Se a linha District Line não estivesse ativa, quais estações (se houvesse alguma) estariam inacessíveis por metrô? (c) Se a linha Jubilee não estivesse ativa, quais estações (se houvesse alguma) estariam inacessíveis por metrô?

6.13 Em uma fotocópia das três combinações comuns de motivos de redes de controle (Quadro, página 355), (a) indique quais nós são controlados apenas por *um* nó acima; (b) indique qual nó exerce controle sobre apenas *um* nó abaixo.

6.14 Em uma fotocópia do fragmento simplificado da rede reguladora de levedura (Figura 6.8), indique exemplos dos motivos da rede de controle (a) forquilha, (b) *"one-two punch"* e (c) adicione uma seta para criar um motivo disperso.

6.15 No dímero formado pela sintropina e pela óxido nítrico-sintase neuronal (Figura 6.9b): (a) A estrutura formada é um dímero aberto ou fechado? (b) Qual elemento de estrutura secundária é compartilhado entre os dois domínios?

6.16 Na rede reguladora transcricional completa de levedura, o número de conexões de entrada nos genes-alvo segue uma distribuição exponencial. Isto é, a probabilidade de um gene ser controlado por k reguladores de transcrição é proporcional a $e^{-\alpha k}$, com $\alpha = 0{,}8$, $k = 1, 2, \ldots$. Qual é a razão entre o número de genes-alvo que recebe quatro conexões de entrada e o número de genes que recebe duas conexões de entrada?

Problemas

6.1 Qual é o número total de posições existentes no microarranjo da Gravura VII? (b) Quantas são complementares a RNAs de fígado? (c) Quantas são complementares a RNAs de cérebro? (d) Quantas são complementares a RNAs de fígado e de cérebro? (e) Quantas não são complementares nem a RNAs de fígado nem a RNAs de cérebro?

6.2 Para avidina-biotina, $K_D = 10^{-15}$. Suponha que o valor de k_{on} seja tão rápido quanto o limite de difusão, $\sim 10^{-9}\,M^{-1}\,s^{-1}$. (a) Qual é o valor de k_{off}? (b) Qual seria o tempo de meia-vida do complexo avidina-biotina? (c) Suponha que o valor de k_{on} para avidina-biotina seja $10^{-7}\,M^{-1}\,s^{-1}$. Qual seria a meia-vida do complexo?

6.3 O fármaco antituberculose isoniazida requer a ativação pela enzima KatG (uma catalase-peroxidase) de *M. tuberculosis*, mas o fármaco relacionado etionamida não exige ativação. Suponha que os perfis de expressão foram medidos nas seguintes situações:

(a) uma cepa com KatG ativa, não exposta a nenhum dos fármacos,

(b) uma cepa com KatG ativa, exposta à isoniazida,

(c) uma cepa sem KatG ativa, exposta à isoniazida,

(d) uma cepa com KatG ativa, exposta à etionamida,

(e) uma cepa sem KatG ativa, exposta à etionamida.

Quais dois perfis entre (b), (c), (d) ou (e) mostrariam um aumento similar de expressão gênica em relação ao perfil (a)? Por que você esperaria padrões de aumento de expressão similares em (b), (d) e (e), mas não em (c)?

6.4 J. Foote e G. Winter compararam as constantes de dissociação do anticorpo natural de camundongos antilisozima (D1.3) com um anticorpo "humanizado" engenhado no qual o sítio de ligação ao antígeno foi inserido em uma estrutura humana (humano-original) e em vários mutantes da "forma humanizada", incluindo humano-mutado. O antígeno era a lisozima da clara de ovo.

Anticorpo	Número de diferenças em relação à seqüência D1.3	k_{on} $M^{-1}s^{-1}$	K_D
D1.3	0	$1{,}4 \times 10^{-6}$	$3{,}7 \times 10^{-9}$
Humano-original	48	$0{,}7 \times 10^{-6}$	260×10^{-9}
Humano-mutado	44	$1{,}3 \times 10^{-6}$	14×10^{-9}

(a) Calcule o k_{off} (taxa de dissociação) para cada anticorpo. (b) O que tem efeito maior na constante de dissociação: diferenças no valor de k_{on} ou de k_{off}?

6.5 Analise o mapa do metrô de Londres contando o número de conexões feitas em cada estação da zona 1 (região central). Conte as conexões *para* estações dentro e fora da zona 1, conquanto sejam originadas na zona 1. Considere uma conexão apenas se duas estações estiverem conectadas por mais de uma linha; em outras palavras, para cada estação a questão é: quantas outras estações podem ser alcançadas sem passar por paradas intermediárias?

(a) Qual é o número máximo de conexões para qualquer estação?

(b) Para cada inteiro k a partir de 1 até o número máximo, quantas estações possuem k conexões?

(c) Represente estes dados graficamente em um gráfico log × log. A relação entre os dados parece linear?

(d) Se a resposta da questão acima for positiva, trace uma linha reta no gráfico log × log e determine o seu expoente.

Resultados de análises de rede deste tipo são mais significativos se os dados cobrirem várias ordens de magnitude, mas isso não é possível neste exemplo.

6.6 Na rede reguladora transcricional completa de levedura, o número de conexões de entrada nos nós tem uma distribuição exponencial. Ou seja, a probabilidade P_k de um gene ser controlado por k reguladores de transcrição é dada por $P_k = Ce^{-\alpha k}$, $k = 1, 2, \ldots$, com $\alpha = 0{,}8$.

(a) Determine a constante de proporcionalidade C em termos de α somando a série $\sum_{k=1}^{\infty} Ce^{-\alpha k} = 1$.

(b) Se $\alpha = 0{,}8$, qual é o valor máximo de k para o qual se esperaria que ao menos 1% dos nós tenha pelo menos k conexões de entrada?

(c) Se $\alpha = 0{,}8$, desenhe o histograma esperado para $1 \leq k \leq 7$.

(d) Determine o valor médio de k em termos de α. (Dica: na solução de (a) você expressou $\sum_{k=1}^{\infty} Ce^{-\alpha k} = 1$ em função de $f(\alpha)$. Tome a derivada desta função com relação a α para chegar à equação $\sum_{k=1}^{\infty} -ke^{-\alpha k} = f'(\alpha)$. Assim, o valor médio de k é dado por $-f'(\alpha)/f(\alpha)$.)

(e) Qual é o valor médio de k ($<k>$) correspondente a $\alpha = 0{,}8$?

(f) Qual é a mediana de k? Este é o valor de κ onde metade dos nós tem um número de conexões de entrada $\leq \kappa$ e metade dos nós tem um número de conexões de entrada $\geq \kappa$. Encontre κ em termos de α. (Dica: se $\sum_{k=1}^{\infty} Ce^{-\alpha k} = 1$, então $\sum_{k=\kappa+1}^{\infty} Ce^{-\alpha k} = \frac{1}{2}$. Mas $\sum_{k=\kappa+1}^{\infty} Ce^{-\alpha k} = e^{-\alpha\kappa}\sum_{k=\kappa+1}^{\infty} Ce^{-\alpha k}$. Em geral, esta abordagem fornecerá uma estimativa não inteira de κ, simplesmente arredonde para o número inteiro mais próximo.)

(g) Se $\alpha = 0{,}8$, qual é a mediana de κ? Como este valor se compara ao valor médio de k ($<k>$)? Estes dois valores são aproximadamente iguais?

6.7 Indique como conectar uma seleção dos três motivos comuns em redes de controle de modo que um único nó de entrada possa influenciar três nós de saída.

Weblemas

6.1 Defina os seguintes termos: (a) *"interactome"*, (b) *"metabolome"* e (c) *"signalome"*. (d) Mais complicado: você pode imaginar e definir um termo "ome" que ainda não tenha sido proposto?

6.2 A catalase-peroxidase KatG de *M. tuberculosis* ativa o fármaco isoniazida. A mutação mais comum da KatG associada com cepas resistentes é S315T. A partir do modelo de *M. tuberculosis*, baseado na estrutura da proteína homóloga de *Burkholderia pseudomallei* [código PDB: 1MWV], sugira um mecanismo para a redução da atividade do mutante na ativação da isoniazida enquanto mantém alguma atividade contra substratos menores.

6.3 Encontre um fragmento da árvore genealógica das famílias reais européias que não seja um grafo.

6.4 Um modelo de crescimento de redes sem escala sugere que se poucos nós e arestas forem adicionados de acordo com a regra de que a probabilidade de adicionar uma nova aresta seja proporcional ao número de arestas que um nó já possui, a rede permanecerá sem escala e manterá o mesmo expoente. Utilizando mapas históricos do metrô de Londres, verifique se as redes antigas eram redes sem escala. Teste se a adição de arestas à rede seguiu a regra de que as arestas foram adicionadas preferencialmente aos nós mais conexos.

6.5 A partir do banco de dados SPiD de interação de proteínas em *B. subtilis* (ver Figura 6.6), (a) que tipo de evidência experimental conecta as proteínas DnaC e DnaG? (b) que tipo de evidência experimental conecta as proteínas DnaC a DnaI?

6.6 A partir do banco de dados SPiD de interação de proteínas em *B. subtillis* (ver Figura 6.6), desenhe um grafo que mostre não apenas os vizinhos imediatos da proteína DnaC, mas também os vizinhos secundários (os vizinhos dos vizinhos).

Conclusões

Como podemos extrapolar da conjuntura atual para a bioinformática do futuro? Claramente, a coleta e o acúmulo de dados prosseguirão e continuarão crescendo. Novas técnicas de produção rápida e em larga escala de dados fornecerão novos tipos de conhecimentos, incluindo informações sobre a integração e o controle dos processos da vida. Recursos computacionais de capacidade crescente serão utilizados no armazenamento, na distribuição e na análise dos resultados. Novos bancos de dados surgirão na *web*, e as conexões entre esses bancos serão mais eficientes. Algoritmos aperfeiçoados serão desenvolvidos para analisar e interpretar as informações fornecidas e transformá-las de puros dados em conhecimento e saber.

Um limiar será atingido quando a nossa compreensão de seqüências e estruturas tornar-se mais completa, no sentido de um subconjunto razoavelmente completo de dados disponíveis sobre as formas vivas atuais ter sido coletado (não deve haver dúvidas de que não somos capazes de conhecer tudo). Isso será reconhecido de maneira prática quando uma investigação aleatória de um genoma ou a determinação de uma nova estrutura protéica mostrar que é mais fácil obter algo já conhecido do que descobrir alguma coisa nova. A natureza é, afinal, um sistema de possibilidades ilimitadas, mas de opções finitas.

Aplicações dos conhecimentos gerados pela bioinformática se tornarão mais factíveis e se desenvolverão mais rapidamente da pesquisa básica à prática industrial e clínica padrão. Alguns dos níveis mais superiores de transferência da informação biológica – tais como os programas de desenvolvimento genético durante a vida dos indivíduos e as atividades da mente humana – serão incluídos nos processos que poderemos descrever quantitativamente e analisar no nível das moléculas e suas interações.

Nas pinturas de Michaelangelo no teto da Capela Sistina, a serpente oferecendo a fruta da árvore do conhecimento para Eva é representada com suas pernas enroladas em volta da árvore na forma de uma dupla hélice. Esperamos que nossos desejos veementes pelo conhecimento contido em outra dupla hélice (DNA) tenham conseqüências mais prósperas.

Respostas aos Exercícios

São fornecidas respostas para todos os exercícios, exceto para aqueles que exigem anotação ou desenho de diagramas.

1.1 (a) 5.000 (b) 5×10^6 (c) Assumindo a população dos EUA como 3×10^8, as demandas de armazenamento para todas as seqüências de DNA de todos os habitantes dos EUA excedem as demandas de armazenamento para EOS/DIS por um fator de 60.

1.2 (a) 20.833 (b) 4,6 (c) 1

1.3 Doença neurodegenerativa, bloco de poliglutamina, huntingtina, prevenção, aconselhamento

1.4 (a) Leu-Ala-His-Lis-Tir

(b) ... cta ...

(c) ... ccg ... (Leu → Pro ...)

(d) ... aag → tag ...

(e) ... taa → tac

1.7 Escolha (a) $E < 10^{-200}$ (b) $E \sim 0{,}003$

1.8 Pelo menos 15 bases de comprimento.

1.9 (a) Cerca de 100.000 gerações de seres humanos. (b) 100.000 gerações de bactérias = 3,8 anos.

1.10 (a) Isoleucina (b) Ácido glutâmico (c) Serina (d) Arginina (e) Lisina (f) Fenilalanina

1.11 (a) Para baixo (b) Para baixo

1.12 Duas vezes.

1.14 m

1.16 75%

1.17 Mudar `([A-Z] [a-z] + [a-z]+)` para `([A-Z] ([a-z]+|\.) [a-z]+)`

1.18 Mudar `$species{$1} = 1;` para `$species{$1}++;`

1.19 CGCAAAAAAGCG ou GCGTTTTTTCGC

2.1 Quase 300.000.

2.2 Se uma outra enzima funcionasse no lugar daquela cujo gene foi nocauteado, poderia não haver nenhum efeito no fenótipo.

2.3 Todas elas.

2.4 Aterrissagem na Lua.

2.5 Autossômico dominante.

2.6 O fenótipo se correlaciona com o alelo do gene do retinoblastoma (RB1), e não com o alelo do gene da esterase D.

2.7 Por causa dos diferentes padrões de expressão nos diferentes tecidos.

2.8 $7{,}5 \times 10^5$ pb, 6 genes.

2.9 O risco não é mais alto do que se o pai fosse normal. (Porém, ver Schwartz, M. & Vissing, J. (2002). Paternal inheritance of mitochondrial DNA, *New Eng. J. Med.* 347, 576-80.)

2.10 Faria uma consulta à seqüência do genoma humano.

2.11 Exceções: trpE e trpD, e trpB e trpA, são dois pares de genes que contribuem para uma enzima. trpD codifica [partes de] duas enzimas.

2.12 Uma inversão ocorreu na região em torno do centrômero.

2.13 As regiões fluorescentes não desapareceriam, mas apareceriam em outra parte.

2.14 A taxa de transferência média é: $5{,}24 \times 10^{-5}$ ORFs/ano, 5×10^{-5} kpb por ano.

2.15 (a) Ambas (b) Um genoma de um ser vivo (c) Ambos (d) Ambos (e) Ambos (f) Bancos de dados computacionais

3.1 Bicicleta, triciclo: propulsão humana.
Bicicleta, motocicleta: número de rodas = 2.
Bicicleta, carro: (número de rodas = 2 AND propulsão humana) OR (número de rodas = 4 AND propulsão a motor).
Triciclo, motocicleta: (número de rodas = 3 AND propulsão humana) OR (número de rodas = 2 AND propulsão a motor).
Triciclo, carro: número de rodas > 2; motocicleta, carro: propulsão a motor.

3.3
```
while(<>){
  if (/\^DEFINITION/) \{s/^DEFINITION\s*//;print \"$_\";\}
  elsif (/^\s*\d+\s*(\[a-z +])\s*$/) {$seq = $1; $seq =~ s/ //g;
  print \"$1\n\";}
}
```

4.1 4

4.2 6

4.3 agtcc → cgtcc → cgtca → cgctca

4.4 (a) Um pouco deslocado da diagonal principal, espere regiões de pareamentos das palavras "time" e "waste" paralelas à diagonal principal. (b) Utilize o programa PERL, na página 182, com janela = 4 e limiar = 2.

4.5 Janela = 2, Limiar = 2

4.6 Com PAM250 H ↔ R é mais provável e com BLOSUM62 W ↔ F é mais provável.

4.7
```
THE.RETORT.COURTEOUS-
THE.REPLY-.CHURLI--SH
```

4.8 Fixar os pesos de todas as rotas para dentro e fora de Uppsala com valores bastante grandes e negativos.

4.9 Faça um gráfico de pontos de uma seqüência contra o complemento reverso da outra.

4.10 Troque a linha

```
$_ =~ /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(.*)\n([A-Za-z\n]*)
    \*\s*\n(.*)\     n([A-Za-z\n]*)\*/;
```

por

```
$_ =~ /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(^\s*\>.*)\
    n([A-Za-z\n]*)\n(^\s*\>.*)\ n([A-Za-z\n]*)\*/;
```

4.11 0,16

4.12 (a) Mais similar (b) Mais similar (c) Pouco similar (d) Menos similar

4.15 4.000.000

4.16 65 = Min(40 + 25, 45 + 20, 50 + 25) onde 40 + 25 corresponde ao movimento vertical, 45 + 20, ao movimento diagonal (malpareado) e 50 + 25 corresponde ao movimento horizontal. As duas setas indicam a possibilidade de dois predecessores com o mesmo valor.

4.18 30, 60, 70, 84

4.19 (a) VDFSAT (b) O escore de VDFSAT = –1.533 e o escore de VDFSAE = –1.503. Escores computados ao se adicionar seis termos, o primeiro dos quais é: o escore do inventário de V × $\sum_{i=1}^{20}$ *BLOSUM*62(V, aminoácido *i*)

4.20 (a)

Número do resíduo	Número de cada aminoácido																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
90					8			2	2	1								3		
91	1			1				2						1		7		4		
92						16														
93	15	1																		
94			2						2	2		6				1		3		
95				4					5				3			2		1		

(b) 32 (c) 49

4.21 (a) Sim (b) Não

4.22 9: a(bc), a(bc), a(cb), ...

4.23 A matriz de distância reduzida após combinar ATCC e ATGC é mostrada no texto. O próximo passo é combinar TTCG e TCGG. Na matriz de distância ainda mais reduzida, após esta combinação, a distância de {ATCC,ATGC} a {TTCG, TCGG} = 3. Portanto, as distâncias da raiz a {ATCC,ATGC} ou a {TTCG, TCGG} = ½ × 3 = 1,5.

4.24 A distância entre ATCC e ATGC = 1, calculada com ambos os métodos, a matriz original e a soma das arestas ao longo do caminho mais curto através da árvore.

A distância entre ATCC e TTCG = 2 quando calculada com a matriz original e 4,5 quando calculada como a soma das arestas ao longo do caminho mais curto através da árvore.

A distância entre ATCC e TCGG = 4 quando calculada com a matriz original e 4,5 quando calculada como a soma das arestas ao longo do caminho mais curto através da árvore.

A distância entre ATGC e TTCG = 3 quando calculada com a matriz original e 4,5 quando calculada como a soma das arestas ao longo do caminho mais curto através da árvore.

A distância entre ATGC e TCGG = 3 quando calculada com a matriz original e 4,5 quando calculada como a soma das arestas ao longo do caminho mais curto através da árvore.

A distância entre TTCG e TCGG = 2, calculada com ambos os métodos, a matriz original e a soma das arestas ao longo do caminho mais curto através da árvore.

4.25 Três nós, com todos os pares de nós conectados.

4.26 Uma ovelha.

4.28 (a) Como n^2 (b) Como n

5.1 25,5 kJ/ligação de hidrogênio.

5.2 (a) Ortólogas (b) Parálogas (c) Parálogas (d) Parálogas (e) Parálogas (f) Nenhum dos dois

5.5 b

5.6 (a) Da terceira à sexta fita a partir da direita no molde, e as hélices que as conectam, possuem a mesma topologia que a segunda até a quinta fita a partir da direita no alvo. A fita mais à direita no alvo corresponde a uma hélice, e a segunda fita mais à direita no molde (que aponta na direção oposta da fita mais à direita no alvo). A fita localizada mais à direita no molde não corresponde a uma fita no alvo. Um motivo $\beta - \alpha - \beta$ extra aparece no alvo, à esquerda da folha beta. (b) As seis fitas na predição de Murzin têm a topologia correta como no alvo; a fita mais à esquerda (N-terminal) no alvo está ausente na predição. (c) A conectividade da fita localizada mais à direita é diferente; o resultado da predição é parecido com a estrutura do alvo, e a estrutura experimental não. A direção da sexta fita a partir da direita é diferente entre o molde e o alvo.

5.8 Você procuraria por proteínas homólogas que tivessem maneiras (separáveis) de se determinar a especificidade. Ou, melhor ainda, você buscaria por proteínas que executassem funções essenciais ao patógeno, mas que não possuíssem homólogas em humanos.

5.9 24

5.10 Ele seleciona pontos acima e à direita da linha $x + 2y = 2$. (Esta linha intercepta o eixo x em $x = 2$ e o eixo y em $y = 1$.)

5.11 O esboço parece com o diagrama na parte superior esquerda da página 265. Interpretação geométrica: selecione pontos abaixo e à direita da linha $y = x$.

6.1 (a) Sim (b) Sim (c) Não

6.3 (a) RVVWLL (b) y5

6.4 (a) 22 (b) A $\leftrightarrow$ C (c) Sim (d) 23 (e) A $\leftrightarrow$ C (f) Sim

6.5 Cerca de 400 seqüências distintas se a recombinação ocorresse com a mesma probabilidade em toda parte. 0,75%

6.6 (a) Exibição por fagos (b) Ressonância plasmônica de superfície

6.7 (a) V_1-V_3
(b) V_1-V_3
(c) V_5-V_6
(d) V_1-V_4
(e) V_4-V_3

6.8 (a) Árvores filogenéticas, partes de vias metabólicas, padrões de citação, a *World Wide Web*. (b) Vias metabólicas, padrões de ligações químicas (se ligações simples/dupla/tripla são distinguíveis).
(c, d)

Grafo	Vértices	Arestas
Grupos de pessoas que se encontraram	pessoas	relações
Sistemas de distribuição de eletricidade	usina elétrica	cabos elétricos
Árvores filogenéticas	espécies (ou taxa)	ramos de descendência
Vias metabólicas	metabólitos	reações catalisadas por enzimas
Padrões de ligações químicas em moléculas	átomos	ligações
Padrões de referências na literatura científica	artigos	referências
A *World Wide Web*	sítios	referências

6.9 (a) 10 (b) 17 (c) 178,5 kJ

6.10 1/21

6.11 (a) Moorgate → Bank → Waterloo → Embankment
(b) King's Cross → Russel Square → Holborn → Tottenham Court Road → Oxford Circus → Euston Square → King's Cross
(c) 2/15

6.12 (a) Amersham – Upminster.
(b) Estações entre Upney e Upminster, estações entre West Brompton e Wimbledon, Chiswick Park (permitindo acesso pela Ferrovia Nacional no entroncamento de Richmond).
(c) Estações entre Stanmore e Baker Street (exceto Wembley Park, West Hampstead – acessíveis pela Ferrovia Nacional – e Finchley Road, Southwark, Bermondsey).

6.15 (a) Fechada (b) Folha β

6.16 0,201

Glossário

alinhamento de seqüências par a par atribuição de correspondências entre resíduos similares em duas seqüências relacionadas

alinhamento múltiplo de seqüências atribuição de correspondências entre resíduos similares em três ou mais seqüências relacionadas

alvo de fármacos uma proteína cuja função poderia ser modificada para aliviar um sintoma de doença

análise de mutação por microarranjos microarranjos projetados para identificar polimorfismos de um único nucleotídeo (SNPs)

árvore um grafo conexo com exatamente um caminho entre cada dois pontos

binomial identificador de uma unidade taxonômica em termos do gênero e da espécie

bootstrapping um protocolo estatístico para validar um método no qual grupos de dados são amostrados aleatoriamente a partir de um conjunto completo de dados, permitindo cópias do mesmo grupo de dados

caminho uma seqüência consecutiva de arestas em um grafo

***chip* de expressão** um microarranjo de DNA contendo oligonucleotídeos derivados de cDNAs

comprimento da aresta um número atribuído a cada aresta de um grafo significando, de certa forma, a distância entre os vértices conectados pela aresta

comprimento do caminho a soma dos comprimentos das arestas em um caminho

deriva gênica alteração nas freqüências alélicas dentro de uma população por outros mecanismos que não a seleção natural

distância de Hamming uma medida da dessemelhança entre duas seqüências de caracteres de mesmo comprimento, igual ao número de caracteres desalinhados em posições correspondentes

distância de Levenshtein uma medida da dessemelhança entre duas seqüências de caracteres, igual ao número mínimo de operações de edição (substituição, inserção, deleção) necessário para converter uma seqüência em outra

docagem molecular predição computacional de possíveis interações e modo de ligação entre ligantes (pequenas moléculas) e proteínas ou ácidos nucléicos

domínio um elemento estrutural compacto de uma proteína que possui estabilidade própria e se enovela independentemente do restante da proteína

duplicação gênica um tipo de mutação em que duas ou mais cópias de um gene parecem corresponder a uma única cópia do gene original

ecossistema um agrupamento de organismos vivos que habitam um local determinado interagindo entre si e com o seu ambiente

efeito hidrofóbico interações desfavoráveis entre solutos apolares com a água, resultando em baixa solubilidade

enzima de restrição uma enzima que cliva o DNA em uma seqüência de bases específica

escore Z uma medida da imprevisão de um resultado em uma distribuição de possíveis resultados; escore Z = (escore – média da distribuição)/desvio-padrão da distribuição

estrutura primária o conjunto de ligações químicas em uma proteína ou ácido nucléico

estrutura quaternária a composição e o agrupamento de subunidades

estrutura secundária estruturas-padrão, estabilizadas por ligações de hidrogênio, entre resíduos em uma proteína ou um ácido nucléico

estrutura supersecundária combinações de elementos de estrutura secundária que são consecutivos na seqüência e interagentes no espaço tridimensional

estrutura terciária o arranjo espacial da cadeia polimérica de uma proteína ou de um ácido nucléico

eucarioto um organismo cujas células contêm um núcleo

evolução convergente o aparecimento de características semelhantes em organismos de origens diferentes e que não descendem de um ancestral comum

éxon segmento expresso de um gene eucariótico que efetivamente codifica parte da seqüência de aminoácidos de uma proteína

expressão regular um padrão dentro de um texto, incluindo caracteres específicos ou classes de caracteres, que pode ser consultado por programas de computador

farmacóforo subestrutura comum a diferentes moléculas que compartilham uma atividade farmacológica

fator B uma medida da precisão com a qual a posição de um átomo é determinada pela cristalografia por difração de raios X

fenótipo as características de um organismo, incluindo as macroscópicas e bioquímicas, exclusivas da seqüência nucleotídica do material genético nas células do organismo

fluxo gênico redistribuição de freqüências alélicas dentro de uma população ou sua transferência entre populações

genética reversa identificação de uma seqüência gênica correspondente a um caráter fenotípico

genômica e proteômica diferencial comparações de seqüências ou padrões de expressão gênica entre indivíduos; por exemplo, entre seres humanos ou animais saudáveis e doentes, ou entre microrganismos infecciosos suscetíveis e resistentes a fármacos

genótipo a informação hereditária contida na seqüência nucleotídica do material genético nas células de um organismo

grafo uma estrutura abstrata contendo vértices (pontos) e arestas (linhas conectando pontos)

grafo conexo um grafo no qual há pelo menos um caminho entre cada dois vértices

hibridização genômica técnica de microarranjos de DNA para detecção de ganhos e perdas de genes ou do seu número de cópias

hibridização *in situ* por fluorescência identificação da posição de uma seqüência de DNA específica em um cromossomo por etiquetagem de um oligonucleotídeo complementar com um marcador fluorescente

híbridos de células somáticas células de roedores contendo um ou uns poucos cromossomos humanos ou fragmentos de cromossomos humanos

homólogo descendente de um ancestral comum

íntron segmento interveniente, inexpressável, de um gene eucariótico que efetivamente não codifica qualquer parte da seqüência de aminoácidos de uma proteína

Ionização por Vaporização Eletrostática (Electrospray Ionization – ESI) um método não-destrutivo para vaporizar e ionizar proteínas, no qual a amostra de proteína é borrifada através de um pequeno tubo capilar com um campo elétrico em sua ponta

jackknifing um protocolo estatístico para validar um método no qual grupos de dados são amostrados aleatoriamente a partir de um conjunto completo de dados

janela o tamanho de cada região consecutiva e sucessiva em uma seqüência, considerada em um tempo particular

LINE (long interspersed nuclear element) elemento nuclear intercalado longo

macroevolução evolução em larga escala em genomas dentro de populações
mapa de contíguos um mapa mostrando a ordem de aparecimento, em uma molécula de DNA, de pequenos fragmentos seqüenciados e com superposição
mapa de restrição um mapa esquemático das posições dos sítios de clivagem para uma ou mais enzimas de restrição
Matrix-Assisted Laser Desorption Ionization (MALDI) Ionização/Dessorção de Matriz Assistida por Laser é um método não-destrutivo de vaporização de proteínas no qual a amostra é fixada a uma mistura compilada de substâncias (a matriz) que absorve inicialmente a energia de um pulso de laser
matriz de substituição uma tabela especificando a probabilidade esperada relativa para trocar qualquer aminoácido (ou nucleotídeo) por outro
microevolução mudanças de genes em pequena escala dentro de uma população, algumas das quais nem sequer afetam o fenótipo
modelagem por homologia predição da estrutura tridimensional de uma cadeia polipeptídica com base na estrutura conhecida adotada por uma seqüência similar
modificação pós-traducional mudança química na cadeia polipeptídica após sua síntese no ribossomo
mutação mudança herdável na seqüência nucleotídica do material genético
PAM (percent accepted mutation) uma medida da dessemelhança entre duas seqüências
percepção de *quorum* comunicação intercelular em uma população de bactérias na qual uma molécula mensageira secretada induz uma mudança fisiológica quando sua concentração excede um valor limite
perfil de hidrofobicidade gráfico da hidrofobicidade de resíduos de aminoácidos sucessivos de uma proteína, fornecendo informações sobre posições de voltas (*turns*) ou regiões encobertas/expostas na proteína
população um grupo de organismos interagentes da mesma espécie
predição da estrutura secundária um esforço para computar, a partir da seqüência de aminoácidos, que regiões da cadeia polipeptídica formam hélices α e fitas de folhas β, sem conhecer a estrutura tridimensional, ou uma predição equivalente das regiões de pareamento de bases de RNA de fita simples
procarioto um organismo cujas células não contêm um núcleo
proteínas modulares proteínas compostas de uma ordenação linear de domínios
Quantitative Structure-Activity Relationships (QSAR) Relações Quantitativas Estrutura-Atividade são correlações entre características estruturais de moléculas e suas reatividades ou efeitos farmacológicos
recombinação um rearranjo do material genético através da permuta de partes dos cromossomos
reconhecimento de padrão de enovelamento esforço para decidir qual (se algum) dos padrões de enovelamento de uma dada biblioteca de enovelamentos é adotado por uma seqüência polipeptídica
redução de dimensionalidade simplificação de conjuntos de dados complexos por meio da extração de um número relativamente pequeno de quantidades (características) desses dados que represente toda ou uma grande parte da sua variação
RFLP (restriction fragment length polymorphism) polimorfismo de comprimentos de fragmentos de restrição é a distribuição de tamanhos de fragmentos produzidos por clivagem do DNA com uma enzima de restrição
seleção natural mecanismo para mudanças na distribuição das características de uma população baseado na reprodução diferencial conduzida pela adaptação
seqüência de caracteres uma seqüência de letras consecutivas obtidas de um alfabeto específico
similaridade semelhança não necessariamente obtida por homologia
SINE (short interspersed nuclear element) elemento nuclear intercalado curto

sítio com seqüência marcada uma pequena região seqüenciada de DNA, variando de 200 a 600 pares de base, que tem uma ocorrência única e localização mapeada em um genoma

SNP (single nucleotide polymorphism) polimorfismo de um único nucleotídeo é quando uma ou mais bases, dentre as possíveis, aparecem em uma determinada posição do genoma

STRP (short tandem repeat polimorphism) ou microssatélite polimorfismo de repetições curtas em *tandem* (dispostas uma atrás da outra) ou microssatélites são segmentos curtos, de apenas 2 a 5 pares de bases, repetidos, porém presentes como muitas cópias consecutivas

VNTR (variable number tandem repeat) ou minissatélite repetição em *tandem* de número variável ou minissatélites são segmentos curtos repetidos com cerca de 10 a 100 pares de base

vocabulário controlado um conjunto de termos fixos e cuidadosamente definidos aos quais as tabelas de atributos em anotações em bancos de dados devem se restringir

voltas regiões nas quais uma cadeia polipeptídica (ou polinucleotídica) muda de direção no espaço tridimensional

zona de incerteza região de 18 a 25% de identidade de seqüência, na qual é difícil decidir, com base na seqüência apenas, se duas proteínas são relacionadas ou não

Danielle

Índice

A

Agrupamento 219, 225
  hierárquico 219
Alelos 25
Algaritmo de Monte Carlo 281
Algoritmo 36
Alinhamento 46 *Gravura III*
Alinhamento *Gravura IV*
Alinhamento de seqüências 178, 46
  múltiplo 179
  par a par 179
  sistema de pontuação em 191
Alinhamento estrutural 253
Alinhamento global 46, 195-196
Alinhamento local 46, 195-196
  algoritmo de Smith-Waterman 195-196
Alinhamento múltiplo de seqüências 206
Alinhavamento (*therading*, método para reconhecimento de enovelamento) 274
Análise das componentes principais (PCA) 318
Análise de algoritmos 36
Anotação de banco de dados 33
Antropologia 123
Apoptose 340
Archaea 42
Área acessível da superfície 343
Aresta (de um grafo) 224
Árvore 224
Árvore da vida 42
Árvore filogenética 223
Aspirina 303

B

Bactérias 42
Banco de dados BLOCKS 193
Banco de dados da protease do HIV 155-156
Banco de Dados de Ácidos Nucléicos (NDB) 151
Banco de Dados de Estruturas Macromoleculares (MSD) 70
Banco de Dados de Modelagem Molecular (MMDB) 70
Banco de Dados de Proteínas (PDB) 21, 70, 148
  worldwide (wwPDB) 148
Banco de dados de seqüências de DNA 142
Banco de Dados Internacional de Imunogenética (IGMT) 156-157
Bancos de dados 28-29
Bancos de dados de estruturas 21
Bancos de dados de seqüências 21, 28-29
Bancos de dados de seqüências de ácidos nucléicos 142
Bancos de dados de seqüências de proteínas 144
Bancos de dados de vias metabólicas 158-159
Biologia de sistemas 28-29, 74, 331
BioMagResBank 151
Bootstrapping 229
BRCA1 172, 319, *Gravura VIII*

C

Cadeias laterais 245
Caixa TATA 101-102
Cálculos da energia conformacional 275
Cambridge Crystallographic Data Centre 151
Caminho em um grafo 224
Caos 339
CATH (Classe/Arquitetura/Topologia/Superfamília homóloga) 155-156
Classificação biológica 41
Clonagem posicional 96
CLUSTAL-W 82, 169-170
Código genético 27
Co-imunoprecipitação 345
Comissão de Enzimas 71
Complexos de proteínas 340
Composto líder 291, 294
Comprimento da aresta 224
Comprimento do caminho 224
Consórcio Internacional para o Seqüenciamento do Genoma Humano 117
Constante de Michaelis 342
Consultando bancos de dados 30-31

Critical Assessment of Fully Automatic Structure Prediction (CAFASP) 262
Critical Assessment of Structure Prediction (CASP) 72, 262, 307
Cromossomo artificial 97
Cromossomo artificial de bactéria 97
Cromossomo artificial de levedura 97

D

DALI (Distance-matrix ALIgnment) 151, 255
Decomposição de valor singular 318
Deriva genética 25
Descoberta de fármacos 289
Descontrole 340
Desnaturação de proteínas 245
Determinação de estrutura por raios X 344
Diagnóstico pré-natal 328
Dinâmica molecular 275
Disperso 355
Dissociação induzida por colisão (CID) 324
Distância de Hamming 191, 230
Distância de Levenshtein 191, 230
Divergência 340
Divergência de estrutura 257
Divergência de função 286
Divergência de seqüência 257
DNA de cloroplasto 108
DNA mitocondrial 108
Docagem molecular 296
Doença de Huntington 76
Domesticação de gado 124
Domínio 62, 131, *Gravura VI*
Domínio de interação 356
Duplicação gênica 25, 111, 131

E

Efeito hidrofóbico 246
Elastase *Gravura IV*
Elemento nuclear intercalado curto (SINE) 50
Elemento nuclear intercalado longo (LINE) 50
Eletroforese em gel de poliacrilamida (PAGE) 91
Enciclopédia de Genes e Genomas de Kyoto (KEGG) 158-159
Engenharia de proteínas 71
Engenharia de *software* 36
Enovelamento de proteínas 28, 240-241, 248
Ensembl 171-172
ENTREZ 144, 161
ENZYME DB 145
Equilíbrio 339
Escala de tempo da historia da Terra 219
Escore *Z* 203
Escore *P* 203
Espectrometria de massa 74, 91, 321
  seqüenciamento de ácidos nucléicos 326
  seqüenciamento de proteínas 324
Esquemas de pontuação, em alinhamento de seqüências 191
Estabilidade de proteínas 245
Estado estacionário 339
Estado nativo de proteínas 28
Estrutura de proteínas 60
  classificação 64, 155-156, 258
  evolução 256
  predição 71, 260
Estrutura do DNA 26, 28, *Gravura I*
Estrutura primária de proteína 61
Estrutura quaternária de proteínas 61
Estrutura secundária 61, 149
  predição 71, 262, 264, 307
Estrutura supersecundária de proteína 62
Estrutura terciária de proteína 61
Estruturas de dados 36
Etiquetas de seqüências expressas (EST) 156-157, 321
Eukarya 42
"Eva" mitocondrial 80
EVA 269
Evolução 24
Evolução convergente 41
Exibição por fagos 345
Expert Protein Analysis System (ExPASy) 170-171

F

Farmacóforo 255, 295-296
Farmacogenômica 76
Fármacos analgésicos 297, 301
Fase de leitura aberta (ORF) 100-101
Fator B 153-154
Fenótipo 24
Fibrose cística 95
Filogenia 218
FISH (hibridização *in situ* por fluorescência) 97, 133, *Gravura II*
Fita 244
Flavodoxina *Gravura V*
Fluxo gênico 25
Folha 62, 149
Formato FASTA 45
Forquilha 355

G

GenBank 21
Gene *eyeless*, em *Drosophila* 54
Gene supressor de tumor *Gravura II*
Genes HOX 238
Genes PAX-6 53
Genética reversa 94, 96
Genoma 88
  bancos de dados 144
  evolução 125
  tamanhos 87, 88
Genoma mínimo 127
Genômica comparativa 130
Genômica diferencial 77, 293
Genômica estrutural 261
Genótipo 24
Gráfico de pontos 180
  e alinhamento 185, 230
Gráficos moleculares 161, 242
Grafo 224, 335
Grupo de genes da hemoglobina 111

H

Haplótipo 327
Hélice 62, 149, 244
Heteroplasmia 124
Heterozigoto 25
Híbridos de células somáticas 97
Híbridos por radiação 97
Homologia 41, 49, 217
Homoplasmia 124
Homozigoto 25

I

Identificação pessoal 124
Imunoprecipitação da cromatina 345
Indexação de bancos de dados 138
Interface proteína-proteína 343
Ionização por Eletropulverização (ESI) 323
Ionização/Dessorção de Matriz Assistida por Laser-Tempo de Vôo (MALDI-TOF), espectrometria de massa 323
Isoniazida 329, *Gravura IX*

J

Jackknifing 229

L

Leucemia *Gravura II*
Ligações químicas cruzadas 345
LINUS 279
Logo de seqüências *Gravura IVb*

M

Mapa de clones contíguos (*contig*) 98
Mapa de ligação 92
Mapa de Ramachandran (ver mapa de Sasisekharan-Ramakrishnan-Ramachandran) 243
Mapa de Sasisekharan-Ramakrishnan-Ramachandran 243
Mapas genéticos 93
Maria Tifóide 337-338
Matriz de pontuação dependente da posição (PSSM) 211
Matriz de substituição 192
Matriz PAM270 193
Matrizes BLOSUM 193
Máxima parcimônia 226
Máxima verossimilhança 227
MEDLINE 159
Metazoários, árvore filogenética dos 43
Metilação do DNA 90
Método da média aritmética não-ponderada de grupo pareado (UPGMA) 226
Métodos cladísticos 226
Metrô de Londres 334
Microarranjo 73, 313
Microscopia eletrônica 358
MODBASE 261
Modelagem molecular, em descoberta de fármacos 295-296
Modelagem por homologia 71, 262, 270
Modelo Oculto de Markov (HMM) 208, 213
Modificação pós-traducional 92
Mutação 25

N

Nocaute 288
Nomenclatura 41
Núcleo de proteínas homólogas 257

O

OCA 153-154
"one-two punch" 355
Online Mendelian Inheritance in Man (OMIM™) 162, 167
Óperon do triptofano (trp) 102
Organela 108
Ortólogas 286

P

Padrão de co-expressão 345-346
Padrão de distribuição filogenética 345-346
Padrão de enovelamento de proteínas 61

Padrões de bandas de cromossomos 92, 109, 134
Parálogas 286
Percent accepted mutation (PAM) 192
Percepção de *quorum* 333
Perfil de expressão 156-157, 313, 329, *Gravuras VII, VIII, IX*
Perfil de hidrofobicidade 249
Perfis 208
  3D 272
PERL 36
Peso para lacunas 194, 201
Pesquisa por seqüências 52
Phobius 216
Piruvato descarboxilase *Gravura VI*
Polimorfismo de repetições curtas em *tandem* (STRP) 98
Polimorfismo de um único nucleotídeo (SNP) 121
População 25
Precisão e seletividade comparadas 52
Predição de função de proteínas 285
Predição de hélice transmembrana 213, 282-283
Primossomo 346-347
Programação dinâmica 196-197
PROSITE 144
Protein Information Resource (PIR) 144, 169-170
Proteína modular 64, 131
Proteoma 72, 74, 88, 91, 156-157
Pseudogene 110, 120
PSI-BLAST 52, 53, 205, 208, 211
Publicação eletrônica 35
PubMed 159

Q

Quimioinformática 294

R

Raiz quadrada da média dos quadrados dos desvios (comumente chamado de desvio médio quadrático) RMSO 253
Raiz, de uma árvore 221
Recombinação 25
Reconhecimento de padrão de enovelamento 71, 262, 272
Recozimento simulado 282
Recuperação de informação 36
Rede de interação de proteínas 344
Rede de interações 332, *Gravura X*
Redes 333
  do tipo "mundo pequeno" 337-338
Redes de recombinação de domínios 345-346
Redes do tipo mundo pequeno 337-338
Redes neurais 266
Redes reguladoras 346-347
Redução de dimensionalidade 317
Relação filogenética 44, 51
Relações quantitativas estrutura-atividade (QSAR) 292
Repetição em *tandem* de número variável (VNTR) 98
Repetições de poliglutamina 76
Resolução, da determinação de estrutura por raios X 154-155
Ressonância Magnética Nuclear (NMR) 154-155, 344
Ressonância plasmônica de superfície 345
Retinoblastoma 132
Roda helicoidal 251
Rosetta 279

S

SCOP (Structural Classification of Proteins) 70, 155-156, 259, 309, *Gravura V*
Seleção de alvos 293
Seleção natural 25
Sequence Retrieval System (SRS) 168
Similaridade 49
Síndrome de Berardinelli-Seip 99
Sítio com seqüência marcada (STS) 100-101
Sobreposição 253
Sobreposição estrutural 253
Structure Explorer 152
Substituições não-sinônimas (Ka) 125
Substituições sinônimas (Ks) 125
SWISS-MODEL 271
SWISS-PROT 144

T

Tabela de características 143
Tabela de expressão gênica 315
Taxonomia 41
Terapia gênica 77
Testes clínicos 290
Tioredoxina *Gravura IV*
Transferência de energia ressonante fluorescente 345-346
Transferência horizontal de genes 128, 133
Transição 192
Transversão 192
Triagem de duplo-híbrido 344
Trigo 108
Tuberculose 328

U

UniProt 144

V

Valor *E* 203
Variação periódica 339
Vocabulário controlado 139

W

World Wide Web 34-35
Worldwide Protein Data Bank (wwPDB) 148

Z

Zona de incerteza 205

BIBLIOTECA BOOKMAN

Linguagem de Programação/TI

Arnold, K., Golsing, J.; Holmes, D. – A Linguagem de Programação Java – 4.ed.

Deitel, H.M.; Deitel, P.J.; Nieto, T. R.;Lin, t.; Sadhu, P.D. – XML: Como Programar

Deitel, Deitel, Nieto & McPhie – Perl: Como Programar

Flanagan, D. – Java: O Guia Essencial - 5.ed.

Flanagan, D. – JavaScript: O Guia Definitivo – 4.ed.

Galuppo, F.; Matheus, V.; Santos, W. – Desenvolvendo com C#

***Herrington, J.P.** – Dicas de PHP

Horstmann, C. – Big Java

Horstmann, C. – Conceitos de Computação com o Essencial de C++ - 3.ed.

Horstmann, C. – Conceitos de Computação com o Essencial de Java - 3.ed.

Hubbard, J. – Coleção Schaum – Programação em C++ – 2.ed.

Hubbard, J. – Coleção Schaum – Programação com Java - 2.ed

Lippman, S. – C#: Um Guia Prático

Lutz, M.; Ascher, D. – Aprendendo Python - 2.ed.

***Robbins, A. & Beebe, N.H.F.** – Classic Shell Scripting

Roman, E.; Ambler, S.W.; Jewell, T. – Dominando o Enterprise JavaBeans – 2.ed.

Siever, E. & Cols. – LINUX: O Guia Essencial – 5.ed.

Stroustrup, B. – A Linguagem de Programação C++ – 3.ed.

Tittel, E. – Coleção Schaum – XML

*Livros em produção no momento da impressão desta obra, mas que muito em breve estarão à disposição dos leitores em língua portuguesa.

Introdução à Bioinformática

Arthur M. Lesk

Em 26 de junho de 2000, a conclusão de um esboço da seqüência do genoma humano mudou as ciências biológicas e médicas para sempre. Ela prometeu explicar como nossos genes nos tornam como somos e como funcionamos. No cerne dessa descoberta fundamental encontra-se uma disciplina que é agora uma das ferramentas mais importantes na construção do conhecimento atual da biologia celular e molecular e do desenvolvimento da qualidade da saúde: a bioinformática.

Escrito por um pioneiro do uso da bioinformática em pesquisa, este livro explica como acessar os bancos de dados de genomas e de proteínas e os tipos de questões biológicas e médicas que esses dados e ferramentas podem responder.

O conteúdo é apresentado de forma extremamente didática e conta com uma variedade de exercícios e problemas. O livro explora, ainda, as ferramentas computacionais da bioinformática, não exigindo conhecimento prévio aprofundado. Recursos interativos também foram incluídos para estimular o aprendizado através da experiência.

O que há de novo nesta edição:

- Novo capítulo sobre proteômica e biologia de sistemas, refletindo os avanços atuais na análise de microarranjos e bancos de dados de genômica e proteômica.
- Cobertura mais ampla da bioinformática aplicada ao planejamento e à descoberta de fármacos.
- Referências ao banco de dados de proteínas (PDB) para todas as estruturas citadas no livro, facilitando o acesso à representação 3D dessas estruturas.
- Cada capítulo apresenta um conjunto de exercícios, cujas respostas são encontradas no final do livro.

ISBN 978-85-363-1104-3

artmed®
EDITORA
RESPEITO PELO CONHECIMENTO

artmed®

www.artmed.com.br